

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2002年12月17日
Date of Application:

出願番号 特願2002-365764
Application Number:

パリ条約による外国への出願に用いる優先権の主張の基礎となる出願の国コードと出願番号

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is

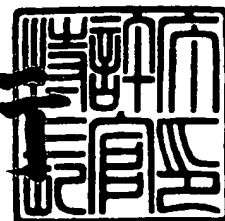
JP2002-365764

出願人 独立行政法人科学技術振興機構
Applicant(s):

2011年 4月12日

特許庁長官
Commissioner,
Japan Patent Office

岩井良行



出証番号 出証特2011-3012903



【書類名】 特許願

【整理番号】 P2234JST

【特記事項】 特許法第 3 0 条第 1 項の規定の適用を受けようとする特
許出願

【提出日】 平成14年12月17日

【あて先】 特許庁長官 殿

【発明者】

 【住所又は居所】 千葉県佐倉市臼井 8 6

 【氏名】 中臺 一博

【発明者】

 【住所又は居所】 東京都渋谷区西原 2 - 1 0 - 9

 【氏名】 奥乃 博

【発明者】

 【住所又は居所】 埼玉県川越市西小仙波町 2 - 1 8 - 3

 【氏名】 北野 宏明

【特許出願人】

 【識別番号】 396020800

 【氏名又は名称】 科学技術振興事業団

【代理人】

 【識別番号】 100082876

 【弁理士】

 【氏名又は名称】 平山 一幸

 【電話番号】 03-3352-1808

【選任した代理人】

 【識別番号】 100069958

 【弁理士】

 【氏名又は名称】 海津 保三

**【手数料の表示】****【予納台帳番号】** 031727**【納付金額】** 21,000円**【提出物件の目録】****【物件名】** 明細書 1**【物件名】** 図面 1**【物件名】** 要約書 1**【包括委任状番号】** 0013677**【プルーフの要否】** 要

【書類名】 明細書

【発明の名称】 ロボット視聴覚システム

【特許請求の範囲】

【請求項 1】 各話者が発した単語とその方向とを組み合わせる複数の音響モデルと、これらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセスを実行する音声認識エンジンと、前記音声認識プロセスによって前記音響モデル別に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結果を選択するセレクトと、を備え、

各話者が同時に発話した単語を各々認識することを特徴とする、ロボット視聴覚システム。

【請求項 2】 前記セレクトが、多数決により前記音声認識プロセス結果を選択するように構成されていることを特徴とする、請求項 1 に記載のロボット視聴覚システム。

【請求項 3】 前記セレクトにて選択された音声認識プロセス結果を外部に出力する対話部を備えていることを特徴とする、請求項 1 又は 2 に記載のロボット視聴覚システム。

【請求項 4】 外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、

上記聴覚イベント、顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてア

ソシエーションストリームを生成するアソシエーションモジュールと、

これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築することにより、音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクトにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項5】 外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位して、ステレオイベントを抽出するステレオモジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモタイイベントを抽出するモータ制御モジュールと、

上記聴覚イベント、顔イベント、ステレオイベント及びモタイイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム、顔ストリーム及びステレオ視覚スト

リームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、

これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクトにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項 6】 前記聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて前記マイクから再び音声を集音させ、この音に対して前記聴覚モジュールにより音源定位・分離された音響信号に基づいて再度聴覚モジュールによる音声認識を行なうように構成されていることを特徴とする、請求項 4 又は 5 に記載のロボット視聴覚システム。

【請求項 7】 前記聴覚モジュールが、音声認識を行なう際に、顔モジュールによる顔イベント又は／及びステレオモジュールによるステレオイベントを参照することを特徴とする、請求項 5 又は 6 に記載のロボット視聴覚システム。

【請求項 8】 前記聴覚モジュールにて判断された音声認識結果を外部に出力する対話部を備えていることを特徴とする、請求項 4～7 の何れかに記載のロボット視聴覚システム。

【請求項 9】 前記アクティブ方向通過型フィルタのパスレンジが、周波数毎に制御可能であることを特徴とする、請求項 4～8 の何れかに記載のロボット

視聴覚システム。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明はロボット、特に人型または動物型ロボットにおける視聴覚システムに関するものである。

【0 0 0 2】

【従来の技術】

近年、このような人型または動物型ロボットにおいては、A I の研究目的の対象にとどまらず、所謂「人間のパートナー」としての将来的な利用が考えられている。そして、ロボットが人間との知的なソーシャルインタラクションを行なうために、視聴覚等の知覚がロボットには必要である。そして、ロボットが人間とのソーシャルインタラクションを実現するためには、知覚のうち、視聴覚、特に聴覚が重要な機能であることは明らかである。従って、視覚、聴覚に関して、所謂能動知覚が注目されてきている。

【0 0 0 3】

ここで、能動知覚とは、ロボット視覚やロボット聴覚等の知覚を担当する知覚装置を知覚すべき目標に追従する働きを言い、例えば、これらの知覚装置を支持する頭部を駆動機構により目標に追従するように姿勢制御するものである。

【0 0 0 4】

ロボットにおける能動視覚においては、少なくとも知覚装置であるカメラが、駆動機構による姿勢制御によって、その光軸方向を目標に向かって保持され、更に目標に対して自動的にフォーカシングやズームイン、ズームアウト等を行う。これにより、目標が移動してもカメラによって撮像される。このような能動視覚の研究が従来、様々に行なわれている。

【0 0 0 5】

これに対して、ロボットにおける能動聴覚においては、少なくとも知覚装置であるマイクが、駆動機構による姿勢制御によってその指向性を目標に向かって保持され、目標からの音がマイクによって集音される。このとき、能動聴覚の不利

な点として、駆動機構が作用している間はマイクが駆動機構の作動音を拾ってしまうために目標からの音に比較的大きなノイズが混入してしまい、目標からの音を認識できなくなってしまうことがある。このような能動聴覚の不利な点を排除するために、例えば視覚情報を参照して音源の方向付けを行なうことにより、目標からの音を正確に認識する方法が採用されている。

【0006】

ところで、このような能動聴覚においては、マイクで集音した音に基づいて、(A) 音源の定位、(B) 各音源から発せられた音毎の分離、(C) そして各音源からの音の認識を行なう必要がある。このうち、(A) 音源定位及び(B) 音源分離については、能動聴覚における実時間・実環境での音源定位・追跡・分離に関する種々の研究が行なわれている（特許文献1参照）。

【0007】

【特許文献1】

国際公開第01/95314号パンフレット

【0008】

ここで、例えば、特許文献1に示すように、HRTF（頭部伝達関数）から求められる両耳間位相差（IPD）、両耳間強度差（IID）を利用して音源定位を行なうことが知られている。また、特許文献1では、例えば所謂方向通過型フィルタ、即ちディレクションパスフィルタを用いて、特定の方向のIPDと同じIPDを有するサブバンドを選択することにより、各音源からの音を分離する方法が知られている。

【0009】

これに対して、音源分離により分離された各音源からの音の認識については、例えばマルチコンディショニングやミッシングデータ等のノイズに対してロバストな音声認識へのアプローチは種々の研究が行なわれている（例えば非特許文献1, 2参照）。

【0010】

【非特許文献1】

J. ベーカー等著、クリーンスピーチモデルに基づくロバスト“ユー

ロスピーチ 2 0 0 1 - 第 7 回ヨーロッパ会議予稿集”，2 0 0 1 年，第 1 巻，
p 2 1 3 - 2 1 6 (J.Baker, M.Cooke, and P.Green, Robust as based on clean
speechmodels: An evaluation of missing data techniques for connected di
git recognition in noise. "7th European conference on Speech Communicatio
n Technology", Volume 1, p. 213-216)

【非特許文献 2】

P. レネベイ等著，ロバストスピーチ認識 ”ユーロスピーチ 2 0 0 1
- 第 7 回ヨーロッパ会議予稿集”，2 0 0 1 年，第 1 2 巻，pp. 1 1 0 7 - 1 1
1 0 (Philippe Renevey, Rolf Vetter, and Jens Kraus. Robust speech recog
nition using missing feature theory and vector quantization. "7th Europe
an Conference on Speech Communication Technology", Volume 12, pp. 1107-1
110)

【0 0 1 1】

【発明が解決しようとする課題】

しかしながら、これらの研究（例えば非特許文献 1，2）においては、S/N
比が小さい場合には、有効な音声認識を行なうことができない。また、実時間・
実環境での音声認識についての研究は行なわれていない。

【0 0 1 2】

この発明は、以上の点に鑑みて、各音源からの分離された音についての認識を
行なうようにしたロボット視聴覚システムを提供することを目的としている。

【0 0 1 3】

【課題を解決するための手段】

上記目的を達成するために、本発明のロボット視聴覚システムの第 1 の構成は
、各話者が発した単語とその方向とを組み合わせる複数の音響モデルと、こ
れらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセ
スを実行する音声認識エンジンと、この音声認識プロセスによって音響モデル別
に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結
果を選択するセレクタと、を備えて、各話者が同時に発話した単語を各々認識す
ることを特徴としている。

【0014】

前記セレクタは、多数決により前記音声認識プロセス結果を選択するように構成され、前記セレクタにて選択された音声認識プロセス結果を外部に出力する対話部を備えていてもよい。

【0015】

このような第1の構成によれば、音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによって、それぞれ音声認識プロセスを行なう。そして、各音響モデルによる音声認識プロセス結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断する。

【0016】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第2の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、上記聴覚イベント、顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または

両耳間強度差（I I D）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクトにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【 0 0 1 7 】

このような第 2 の構成によれば、聴覚モジュールがマイクが集音した外部の対象からの音から、調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て個々の話者を同定し、その聴覚イベントを抽出する。

【 0 0 1 8 】

また、顔モジュールが、カメラにより撮像された画像から、パターン認識による各話者の顔識別と定位から、個々の話者の顔イベントを抽出する。

【 0 0 1 9 】

さらに、モータ制御モジュールが、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによって、モータイベントを抽出する。

【 0 0 2 0 】

なお、上記イベントとは、各時点において検出される音または顔が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するように接続したイベントを示している。

【 0 0 2 1 】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント、顔イベント及びモータイベントに基づいて、各話者の聴覚ストリーム及び顔ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成して、アテンション制御モジュールがこれらのストリームに基づいてアテンション制御を行なうことにより、モータ制御モジュールの駆動モータ制御のプランニングを行なう。なお、アソシエーションストリームとは、聴覚ストリーム及び顔ストリームを包含する概念である。

【 0 0 2 2 】

なお、アテンションとは、ロボットが対象である話者を聴覚的及び／または視覚的に「注目」することであり、アテンション制御とは、モータ制御モジュールによりその向きを変えることによってロボットが上記話者に注目するようにすることである。

【 0 0 2 3 】

そして、アテンション制御モジュールは、このプランニングに基づいて、モータ制御モジュールの駆動モータを制御することにより、ロボットの方向を対象である話者に向ける。これにより、ロボットが対象である話者に対して正対することにより、聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクで正確に集音、定位することができると共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【 0 0 2 4 】

従って、このような聴覚モジュール、顔モジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても、各話者をそれぞれ知覚することができる。

【 0 0 2 5 】

また、例えば聴覚イベントまたは顔イベントの何れか一方が欠落したときであっても、顔イベントまたは聴覚イベントのみに基づいて、対象である話者をアソシエーションモジュールが知覚することができるので、リアルタイムにモータ制御モジュールの制御を行なうことができる。

【 0 0 2 6 】

さらに、上記聴覚モジュールが、上述したように音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクトにより統合して、最も信頼性の高い音声認識結果を判断する。

【 0 0 2 7 】

これにより、従来の音声認識と比較して複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセクタにより統合して、最も信頼性の高い音声認識結果を判断して、より一層正確な音声認識を行なうことができる。

【0028】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第3の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいてピッチ抽出、調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定しその聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えこのカメラで撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出するステレオモジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモタイイベントを抽出するモータ制御モジュールと、前記聴覚イベント、顔イベント、ステレオイベント及びモタイイベントから聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定しこの決定に対してカルマンフィルタを用いて前記イベントを時間方向に接続することにより聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを生成しさらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、音声認識の際に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルに

よる音声認識結果をセクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【0029】

このような第3の構成によれば、聴覚モジュールは、マイクが集音した外部の目標からの音から調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て、個々の話者の方向を決定してその聴覚イベントを抽出する。

【0030】

また、顔モジュールは、カメラにより撮像された画像からパターン認識による各話者の顔識別と定位から各話者を同定して、個々の話者の顔イベントを抽出する。さらに、ステレオモジュールは、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出する。

【0031】

さらに、モータ制御モジュールは、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによってモタイイベントを抽出する。

【0032】

なお、上記イベントとは、各時点において検出される音、顔及び縦に長い物体が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するように接続したイベントを示している。

【0033】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント、顔イベント、ステレオイベント及びモタイイベントに基づいて、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報によって各話者の方向を決定することにより、各話者の聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成する。なお、アソシエーションストリームとは、聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームを包含する概念である。この

際、アソシエーションモジュールは、聴覚イベントの音源定位及び顔イベントの顔定位、即ち聴覚及び視覚の方向情報に基づいて各話者の方向を決定し、決定された各話者の方向を参考にして、アソシエーションストリームを生成する。

【 0 0 3 4 】

そして、アテンション制御モジュールが、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行なう。そして、アテンション制御モジュールは、このプランニングに基づいてモータ制御モジュールの駆動モータを制御してロボットの方向を目標である話者に向ける。これにより、ロボットが目標である話者に対して正対することによって聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクにより正確に集音、定位することができる共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【 0 0 3 5 】

従って、このような聴覚モジュール、顔モジュール、ステレオモジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、聴覚ストリームの音源定位及び顔ストリームの話者定位という方向情報に基づいて各話者の方向を決定することにより、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロボスタ性が向上し、複数の話者であっても各話者をそれぞれ確実に知覚することができる。

【 0 0 3 6 】

また、例えば聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームの何れかが欠落したときであっても、残りのストリームに基づいて目標である話者をアテンション制御モジュールが追跡することができるので、正確に目標の方向を把握して、モータ制御モジュールの制御を行なうことができる。

【 0 0 3 7 】

ここで、聴覚モジュールが、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームやステレオモジュールからのステレオ視覚ストリームをも考慮して音源定位を行なうこ

とによって、より一層正確な音源定位を行なうことができる。

【0038】

そして、上記聴覚モジュールは、アソシエーションモジュールからの正確な音源方向情報に基づいて、聴覚特性に従って正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（IPD）または両耳間強度差（IID）をもったサブバンドを集めて、音源の波形を再構築して音源分離を行なうので、上述した聴覚特性に応じてパスレンジ即ち感度を調整することにより、方向による感度の違いを考慮して、より正確に音源分離を行なうことができる。さらに、上記聴覚モジュールは、上述したように聴覚モジュールによって音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクトにより統合して、最も信頼性の高い音声認識結果を判断して、この音声認識結果を対応する話者と関連付けて出力する。

【0039】

これにより、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセレクトにより統合して、最も信頼性の高い音声認識結果を判断することにより、より一層正確な音声認識を行なうことができる。

【0040】

なお、第2の構成と第3の構成においては、聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて、前記マイクから再び音声を集音させ、この音に対して聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なうように構成されている。

【0041】

さらに、前記聴覚モジュールは、音声認識を行なう際に顔モジュールによる顔イベントを参照するのが望ましい。また、前記聴覚モジュールにて判断された音

声認識結果を外部に出力する対話部が備えられていてもよい。さらに、前記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であることが望ましい。

【 0 0 4 2 】

上記聴覚モジュールによる音声認識ができなかったとき、アテンション制御モジュールが、当該音響信号の音源の方向（当該話者）にマイク及びカメラを向けて、再度マイクから音声を集音させ、聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なう場合には、ロボットの聴覚モジュールのマイク及び顔モジュールのカメラが当該話者と正対することによって、確実な音声認識を行なうことが可能になる。

【 0 0 4 3 】

上記聴覚モジュールは、音声認識を行なう際に、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームをも考慮する。即ち、聴覚モジュールは、顔モジュールにより定位された顔イベントに関して、聴覚モジュールにより定位・分離された音源（話者）からの音響信号に基づいて音声認識を行なうことにより、より一層正確な音声認識を行なうことができる。

【 0 0 4 4 】

上記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であると、さらに集音した音からの分離の精度が上がり、これにより音声認識もさらに向上する。

【 0 0 4 5 】

【発明の実施の形態】

以下、図面に示した実施形態に基づいて、この発明を詳細に説明する。

図 1 及び図 2 は、それぞれこの発明によるロボット視聴覚システムの一実施形態を備えた実験用の上半身のみの人型ロボットの全体構成例を示している。図 1 において、人型ロボット 10 は、4 D O F（自由度）のロボットとして構成されており、ベース 11 と、ベース 11 上にて一軸（垂直軸）周りに回動可能に支持された胴体部 12 と、胴体部 12 上にて三軸方向（垂直軸、左右方向の水平軸及

び前後方向の水平軸)の周りに揺動可能に支持された頭部 1 3 とを含んでいる。

【0 0 4 6】

上記ベース 1 1 は固定配置されていてもよく、脚部として動作可能としてもよい。また、ベース 1 1 は、移動可能な台車等の上に載置されていてもよい。胴体部 1 2 は、ベース 1 1 に対して垂直軸の周りに、図 1 にて矢印 A で示すように回転可能に支持されており、図示しない駆動手段によって回転駆動されると共に、図示の場合、防音性の外装によって覆われている。

【0 0 4 7】

頭部 1 3 は胴体部 1 2 に対して連結部材 1 3 a を介して支持されており、この連結部材 1 3 a に対して前後方向の水平軸の周りに、図 1 にて矢印 B で示すように揺動可能に、また左右方向の水平軸の周りに、図 2 にて矢印 C で示すように揺動可能に支持されていると共に、上記連結部材 1 3 a が、胴体部 1 2 に対してさらに前後方向の水平軸の周りに、図 1 にて矢印 D で示すように揺動可能に支持されており、それぞれ図示しない駆動手段によって、各矢印 A, B, C, D 方向に回転駆動される。ここで、頭部 1 3 は、図 3 に示すように全体が防音性の外装 1 4 により覆われ、前側にロボット視覚を担当する視覚装置としてのカメラ 1 5、両側にロボット聴覚を担当する聴覚装置としての一对のマイク 1 6 (1 6 a, 1 6 b) を備えている。なお、マイク 1 6 は、頭部 1 3 の両側に限定されることなく、頭部 1 3 の他の位置あるいは胴体部 1 2 等に設けられていてもよい。

【0 0 4 8】

上記外装 1 4 は、例えばウレタン樹脂等の吸音性の合成樹脂から構成されており、頭部 1 3 の内部がほぼ完全に密閉されて、頭部 1 3 の内部の遮音が行われるように構成されている。なお、胴体部 1 2 の外装も、頭部 1 3 の外装 1 4 と同様に、吸音性の合成樹脂から構成されている。

【0 0 4 9】

上記カメラ 1 5 は公知の構成であって、例えば所謂パン、チルト、ズームの 3 DOF (自由度) を有する市販のカメラにより構成されている。尚、カメラ 1 5 は、同期をとってステレオ画像を送ることができるように設計されている。

【0 0 5 0】

上記マイク 16 は、それぞれ頭部 13 の側面において前方に向かって指向性を有するように取り付けられている。マイク 16 の左右の各マイク 16 a, 16 b は、それぞれ図 1 及び図 2 に示すように、頭部 13 の外装 14 の両側に配置された段部 14 a, 14 b の内側に取り付けられている。そして、各マイク 16 a, 16 b は、段部 14 a, 14 b に設けられた貫通穴を通して、前方の音を集音すると共に、外装 14 の内部の音を拾わないように、適宜の手段により遮音されている。なお、段部 14 a, 14 b に設けられた貫通穴は、段部 14 a, 14 b の内側から頭部前方に向けて貫通するように、各段部 14 a, 14 b に形成されている。これにより、各マイク 16 a, 16 b は、所謂バイノーラルマイクとして構成されている。なお、マイク 16 a, 16 b の取付位置に近接する外装 14 は人間の外耳形状に形成されていてもよい。ここで、マイク 16 は、外装 14 の内側に配置された一对の内部マイクを含んでいてもよく、この内部マイクにより集音された内部音に基づいて、ロボット 10 の内部に発生するノイズをキャンセルすることができる。

【0051】

図 4 は、上記カメラ 15 及びマイク 16 を含むロボット視聴覚の電氣的構成例を示している。図 4 において、ロボット視聴覚システム 17 は、聴覚モジュール 20, 顔モジュール 30, ステレオモジュール 37, モータ制御モジュール 40 及びアソシエーションモジュール 50 から構成されている。

【0052】

ここで、アソシエーションモジュール 50 はクライアントからの依頼に応じて処理を実行するサーバとして構成されており、このサーバに対するクライアントが、他のモジュール、即ち聴覚モジュール 20, 顔モジュール 30, ステレオモジュール 37, モータ制御モジュール 40 であり、これらのサーバとクライアントとは、互いに非同期で動作する。なお、上記サーバと各クライアントとは、各々、パーソナルコンピュータにより構成されており、更にこれらの各パーソナルコンピュータは、例えば TCP/IP プロトコルの通信環境の下で、相互に LAN (Local Area Network) として構成されている。この場合、好ましくは、データ量の大きいイベントやストリームの通信のためには、ギガビット (Giga bit) のデ

ータ交換が可能な高速ネットワークをロボット視聴覚システム 17 に適用するのが好ましく、また時刻の同期等の制御用通信のためには中速ネットワークをロボット視聴覚システム 17 に適用するのが好ましい。このように大きなデータが高速に各パーソナルコンピュータ間を伝送することで、ロボット全体のリアルタイム性及びスケーラビリティを向上させることができる。

【0053】

また、各モジュール 20, 30, 37, 40, 50 は、それぞれ階層的に分散して構成されており、具体的には下位から順次にデバイス層、プロセス層、特徴層、イベント層から構成されている。

【0054】

上記聴覚モジュール 20 は、デバイス層としてのマイク 16 と、プロセス層としてのピーク抽出部 21, 音源定位部 22, 音源分離部 23 及びアクティブ方向通過型フィルタ 23a と、特徴層（データ）としてのピッチ 24, 音源水平方向 25 と、イベント層としての聴覚イベント生成部 26 と、さらにプロセス層としての音声認識部 27 及び会話部 28 と、から構成されている。

【0055】

ここで、聴覚モジュール 20 は、図 5 に示すように作用する。即ち、図 5 において、聴覚モジュール 20 は、例えば 48 kHz, 16 ビットでサンプリングされたマイク 16 からの音響信号を、符号 X1 で示すように FFT（高速フーリエ変換）により周波数解析して、符号 X2 で示すように左右のチャンネル毎にスペクトルを生成する。そして、聴覚モジュール 20 は、ピーク抽出部 21 により左右のチャンネル毎に一連のピークを抽出して、左右のチャンネルで同じか類似のピークをペアとする。

【0056】

ここで、ピーク抽出は、 (α) パワーがしきい値以上で且つ (β) ローカルピークであって、 (γ) 低周波ノイズとパワーの小さい高周波帯域をカットするため例えば 90 Hz 乃至 3 kHz の間の周波数であるという、3 つの条件 $(\alpha \sim \gamma)$ を満たすデータのみを透過させる帯域フィルタを使用して行なわれる。このしきい値は、周囲の暗騒音を計測して、さらに感度パラメータ、例えば 10 dB を

加えた値として定義される。

【0057】

そして、聴覚モジュール20は、各ピークが調波構造を有していることを利用して音源分離を行う。具体的には、音源分離部23は、周波数の低い方から順に調波構造を有するローカルピークを抽出して、この抽出されたピークの集合を一つの音とみなす。このようにして、音源毎の音響信号が混合音からそれぞれ分離される。音源分離の際、聴覚モジュール20の音源定位部22は、符号X3で示すように、音源分離部23にて分離された各音源毎の音響信号に対して、左右のチャンネルから同じ周波数の音響信号を選択して、IPD（相互位相差）及びIID（相互強度差）を計算する。なお、この計算は、例えば5度毎に行われる。そして、音源定位部22は、計算結果をアクティブ方向通過型フィルタ23aに出力する。

【0058】

これに対して、アクティブ方向通過型フィルタ23aは、アソシエーションモジュール50にて算出されたアソシエーションストリーム59の方向 θ に基づいて、符号X4で示すように、IPDの理論値 $IPD(=\Delta\phi'(\theta))$ を生成すると共に、IIDの理論値 $IID(=\Delta\rho'(\theta))$ を計算する。なお、方向 θ は、顔定位（顔イベント39）とステレオ視覚（ステレオ視覚イベント39a）と音源定位（聴覚イベント29）とに基づいて、アソシエーションモジュール50におけるリアルタイムトラッキング（符号X3'）による算出結果である。

【0059】

ここで、理論値IPDと理論値IIDの各計算は、以下に説明する聴覚エビポラ幾何を利用して行われ、具体的にはロボット10の正面を0度と設定し、±90度の範囲で理論値IPD及び理論値IIDが計算される。ここで、上記聴覚エビポラ幾何は、HRTFを使用せずに音源の方向情報を得るために必要である。ステレオ視覚研究においては、エビポラ幾何が最も一般的な定位法の一つであり、聴覚エビポラ幾何は視覚におけるエビポラ幾何の聴覚への応用である。そして、聴覚エビポラ幾何が幾何学的関係を利用して方向情報を得るので、HRTFを不要にすることができるのである。

【0060】

上記聴覚エピソード幾何においては、音源が無限遠にあると仮定し、 $\Delta\phi$, θ , f , v をそれぞれIPD, 音源方向, 周波数, 音速とし、 r をロボット頭部を球形とみなした場合の半径とすると、以下の式(1)

【数1】

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta)$$

により表わされる。

【0061】

他方、FFT(高速フーリエ変換)により得られた一対のスペクトルに基づいて、各サブバンドのIPD $\Delta\phi'$ 及びIID $\Delta\rho'$ を、以下の式(2), (3)により計算する。

【数2】

$$\Delta\phi' = \arctan\left(\frac{\Im[S_{p_l}]}{\Re[S_{p_l}]}\right) - \arctan\left(\frac{\Im[S_{p_r}]}{\Re[S_{p_r}]}\right)$$

【数3】

$$\Delta\rho' = 20\log_{10}\left(\frac{|S_{p_l}|}{|S_{p_r}|}\right)$$

ここで、 S_{p_l} , S_{p_r} は、それぞれある時刻に左右のマイク16a, 16bから得られたスペクトルである。

【0062】

さらに、アクティブ方向通過型フィルタ23aは、符号X7で示す通過帯域関数に従って、前記ストリーム方向 θ_S から、 θ_S に対応するアクティブ方向通過型フィルタ23aの通過帯域 $\delta(\theta_S)$ を選択する。ここで、通過帯域関数は、図5のX7に示すように、ロボットの正面方向($\theta=0$ 度)で感度が最大となり、側方で感度が低下することから、 $\theta=0$ 度で最小値をとり、側方でより大きくなるような関数である。これは、正面方向で定位の感度が最大になり、左右に角度が大きくなるにつれて感度が低下するという聴覚特性を再現するためのものである。なお、正面方向で定位の感度が最大になることは、哺乳類の目の構造に見

られる中心窩にならって聴覚中心窩と呼ぶ。この聴覚中心窩に関して、人間の場合には、正面の定位の感度が ± 2 度程度であり、左右90度付近にて ± 8 度程度とされている。

【0063】

そして、アクティブ方向通過型フィルタ23a'は、選択した通過帯域 $\delta(\theta_S)$ を使用して、 θ_L から θ_H の範囲にある音響信号を抽出する。尚、 $\theta_L = \theta_S - \delta(\theta_S)$ 、 $\theta_H = \theta_S + \delta(\theta_S)$ と定義する。

【0064】

また、アクティブ方向通過型フィルタ23aは、符号X5で示すように、ストリーム方向 θ_S を頭部伝達関数(HRTF)に利用して、 θ_L 及び θ_H におけるIPD及びIIDの理論値 $IPD(=\Delta\phi_H(\theta_S))$ と $IID(=\Delta\rho_H(\theta_S))$ とを、即ち抽出すべき音源の方向を推定する。そして、アクティブ方向通過型フィルタ23aは、音源方向 θ に対して聴覚エビポーラ幾何に基づいて各サブバンド毎に計算された $IPD(=\Delta\phi_E(\theta))$ 及び $IID(=\Delta\rho_E(\theta))$ と、HRTFに基づいて得られた $IPD(=\Delta\phi_H(\theta))$ 及び $IID(=\Delta\rho_H(\theta))$ とに基づいて、符号X6で示すように、前述した通過帯域 $\delta(\theta)$ により決定される角度 θ_L から θ_H の角度範囲で、抽出された $IPD(=\Delta\phi_E)$ 及び $IID(=\Delta\rho_E)$ が以下の条件を満たすようなサブバンドを集める。

【0065】

ここで、周波数 f_{th} は、フィルタリングの判断基準としてIPDまたはIIDを採用する閾値であって、IPDによる定位が有効である周波数の上限を示す。なお、周波数 f_{th} は、ロボット10のマイク間距離に依存し、本実施形態においては例えば1500Hz程度である。

【0066】

即ち、

【数4】

$$\begin{aligned} f < f_{th} & : \quad \Delta\phi_E(\theta_l) \leq \Delta\phi' \leq \Delta\phi_E(\theta_h) \\ f \geq f_{th} & : \quad \Delta\rho_H(\theta_l) \leq \Delta\rho' \leq \Delta\rho_H(\theta_h) \end{aligned}$$

【0067】

これは、所定周波数 f_{th} 未満の周波数で、HRTFによるIPDの通過帯域 $\delta(\theta)$ の範囲内にIPD ($=\Delta\phi'$) が在る場合、そして所定周波数 f_{th} 以上の周波数でHRTFによるIIDの通過帯域 $\delta(\theta)$ の範囲内にIID ($=\Delta\rho'$) が在る場合に、サブバンドを集めることを意味している。ここで、一般に低周波数帯域ではIPDが大きく影響し、高周波数帯域ではIIDが大きく影響し、その閾値である周波数 f_{th} はマイク間距離に依存する。

【0068】

そして、アクティブ方向通過型フィルタ23aは、このようにして集めたサブバンドから音響信号を再合成して、波形を構築することにより、符号X8で示すように、パス-サブバンド方向を生成し、符号X9で示すように、各サブバンド毎にフィルタリングを行なって、符号X10で示す逆周波数変換IFFT（逆フーリエ変換）により、符号X11で示すように、該当範囲にある各音源からの分離音（音響信号）を抽出する。

【0069】

上記音声認識部27は、図5に示すように、自声抑制部27aと自動認識部27bとから構成されている。自声抑制部27aは、聴覚モジュール20にて音源定位・音源分離された各音響信号から、後述する対話部28のスピーカ28cから発せられた音声除去して外部からの音響信号のみを取り出すものである。自動認識部27bは、図6に示すように、音声認識エンジン27cと音響モデル27dとセレクタ27eとから構成されており、この音声認識エンジン27cとしては、例えば京都大学で開発された「Jurian」という音声認識エンジンを利用することができ、これにより各話者が発話した単語を認識することができるようになっている。

【0070】

図6において、自動認識部27bは、例として男性2人（話者A，C）と女性1人（話者B）の三人の話者の認識を行なうように構成されている。このために自動認識部27bには、各話者の各方向毎にそれぞれ音響モデル27dが備えられている。図6の場合には、音響モデル27dは、3人の各話者A，B，Cに関してそれぞれ各話者が発した音声とその方向とを組み合わせで成り、複数種類、

この場合 9 種類の音響モデル 27 d が備えられている。

【0071】

音声認識エンジン 27 c は、並列に 9 つの音声認識プロセスを実行し、その際に上記 9 つの音響モデル 27 d が用いられる。具体的には、音声認識エンジン 27 c は、それぞれ互いに並列的に入力された音響信号に対して、上記 9 つの音響モデル 27 d を用いて音声認識プロセスを実行する。そして、これらの音声認識結果がセレクト 27 e に出力される。

【0072】

上記セレクト 27 e は、各音響モデル 27 d からのすべての音声認識プロセス結果を統合して、例えば多数決により最も信頼性が高い音声認識プロセス結果を判断して、その音声認識結果を出力する。

【0073】

ここで、特定話者の音響モデル 27 d に対する単語認識率を具体的な実験により説明する。

まず、3 m×3 m の部屋内において、3 つのスピーカをロボット 10 から 1 m の位置に且つロボットから 0 度及び±60 度の方向に置く。次に、音響モデル用の音声データとして、男性 2 名、女性 1 名が各々発話した、色、数字、食べ物のような 150 語の単語の音声をスピーカから出力して、ロボット 10 のマイク 16 a, 16 b で集音する。なお、各単語の集音に当たり、一つのスピーカのみからの音声、二つのスピーカから同時に出力される音声、そして三つのスピーカから同時出力される音声、として、各単語に対して 3 つのパターンを録音する。そして、録音した音声信号に対して前述したアクティブ方向通過型フィルタ 23 a によって音声分離して各音声データを抽出し、話者及び方向毎に整理して、音響モデルのトレーニングセットを作成する。

【0074】

そして、各音響モデル 27 d には、トライフォンを使用して、各トレーニングセット毎に、HTK (Hidden Markov Model) ツールキット 27 f を使用して、各話者の各方向毎に計 9 種類の音声認識用の音声データを作成した。

【0075】

このようにして得られた音響モデル用音声データを使用して、特定話者の音響モデル 27 d に対する単語認識率を実験により調べたところ、図 7 に示す結果が得られた。図 7 は、横軸に方向を、縦軸に単語認識率を示すグラフであり、符号 P は本人（話者 A）の音声、符号 Q は他者（話者 B, C）の音声の場合を示す。話者 A の音響モデルでは、話者 A がロボット 10 の正面に位置している場合（図 7（A））には、正面（0 度）にて 80 % 以上の単語認識率となり、また話者 A が右方 60 度または左方 -60 度に位置する場合、それぞれ図 7（B）又は図 7（C）に示すように、話者よりも方向の違いによる認識率の低下が少なく、特に話者も方向もあっている場合には、80 % 以上の単語認識率となることが分かった。

【0076】

この結果を考慮して、音声認識の際に、音源方向が既知であることを利用して、セクタ 27 e は、以下の式（5）により与えられるコスト関数 $V(p_e)$ を統合のために使用する。

【数 5】

$$V(p_e) = \left(\sum_d r(p_e, d) \cdot v(p_e, d) + \sum_d r(p, d_e) \cdot v(p, d_e) - r(p_e, d_e) \right) \cdot P_v(p_e)$$

$$v(p, d) = \begin{cases} 1 & \text{if } Res(p, d) = Res(p_e, d_e) \\ 0 & \text{if } Res(p, d) \neq Res(p_e, d_e) \end{cases}$$

【0077】

ここで、 $r(p, d)$ 、 $Res(p, d)$ をそれぞれ話者 p と方向 d の音響モデルを使用した場合の単語認識率と入力音声に対する認識結果と定義し、 d_e をリアルタイムトラッキングによる音源方向とし、さらに p_e を評価対象の話者とする。

【0078】

上記 $P_v(p_e, d_e)$ は顔認識モジュールで生成される確率であり、顔認識ができない場合には常に 1.0 とする。そして、セクタ 27 e は最も大きいコスト関数 $V(p_e)$ を有する話者 p_e と認識結果 $Res(p, d)$ を出力する。

その際、セクタ 27 e は、顔モジュール 30 からの顔認識による顔イベント 39 を参照することにより、話者を特定することができるので、音声認識のロバスト性を向上させることができる。

【0079】

なお、コスト関数 $V(p_e)$ の最大値が 1.0 以下または二番目に大きい値と近い場合には、音声認識が失敗または一つの候補に絞りきれなかったことにより音声認識ができないと判断して、その旨を後述する対話部 28 に出力する。上記対話部 28 は、対話制御部 28 a と音声合成部 28 b とスピーカ 28 c とから構成されている。上記対話制御部 28 a は、後述するアソシエーションモジュール 60 により制御されることにより、音声認識部 27 からの音声認識結果、即ち話者 p_e と認識結果 $R_{es}(p, d)$ とに基づいて、対象とする話者に対する音声データを生成し、音声合成部 28 b に出力する。上記音声合成部 28 b は、対話制御部 28 a からの音声データに基づいてスピーカ 28 c を駆動して、音声データに対応する音声を発する。

【0080】

これにより、対話部 28 は音声認識部 27 からの音声認識結果に基づいて、例えば話者 A が好きな数字として「1」と言った場合に、ロボット 10 が当該話者 A に正対した状態で、当該話者 A に対して「Aさんは「1」と言いました」というように音声を発することになる。

【0081】

なお、対話部 28 は、音声認識部 27 から音声認識ができなかった旨が出力された場合には、ロボット 10 が当該話者 A に正対した状態で、当該話者 A に対して、「あなたは「2ですか？ 4ですか？」と質問して、再度話者 A の回答について音声認識を行なうようになっている。この場合、話者 A に対してロボット 10 が正対していることから、音声認識の精度がより一層向上することになる。

【0082】

このようにして、聴覚モジュール 20 は、マイク 16 からの音響信号に基づいて、ピッチ抽出、音源の分離及び定位から少なくとも一人の話者を特定（話者同定）してその聴覚イベントを抽出し、ネットワークを介してアソシエーションモ

ジュール 50 に対して送信すると共に、各話者の音声認識を行なって対話部 28 により音声認識結果を話者に対して音声により確認するようになっている。

【0083】

ここで、実際には、音源方向 θ_s が時間 t の関数であることから、特定音源を抽出し続けるためには時間方向の連続性を考慮する必要があるが、上述したように、リアルタイムトラッキングからのストリーム方向 θ_s により、音源方向を得るようにしている。

【0084】

これによって、リアルタイムトラッキングにて、すべてのイベントをストリームという時間的流れを考慮した表現で表わしているので、同時に複数の音源が存在したり、音源やロボット自身が移動する場合でも、一つのストリームに注目することによって特定音源からの方向情報を連続的に得ることができる。さらに、ストリームは視聴覚のイベントを統合するためにも使用しているので、顔イベントを参照して聴覚イベントにより音源定位を行なうことにより、音源定位の精度が向上することになる。

【0085】

上記顔モジュール 30 は、デバイス層としてのカメラ 15 と、プロセス層としての顔発見部 31、顔識別部 32、顔定位部 33 と、特徴層（データ）としての顔 ID 34、顔方向 35 と、イベント層としての顔イベント生成部 36 と、から構成されている。

【0086】

これにより、顔モジュール 30 は、カメラ 15 からの画像信号に基づいて、顔発見部 31 により例えば肌色抽出により各話者の顔を検出し、顔識別部 32 にて前もって登録されている顔データベース 38 により検索して、一致した顔があった場合、その顔 ID 34 を決定して当該顔を識別すると共に、顔定位部 33 により当該顔方向 35 を決定（定位）する。

【0087】

ここで、顔モジュール 30 は、顔発見部 31 が画像信号から複数の顔を見つけた場合、各顔について上記処理、即ち識別及び定位そして追跡を行なう。その際

、顔発見部 31 により検出された顔の大きさ、方向及び明るさがしばしば変化するので、顔発見部 31 は顔領域検出を行なって、肌色抽出と相関演算に基づくパターンマッチングの組合せによって 200 m 秒以内に複数の顔を正確に検出できるようになっている。

【0088】

顔定位部 33 は、二次元の画像平面における顔位置を三次元空間に変換し、三次元空間における顔位置を、方位角 θ 、高さ ϕ 及び距離 r のセットとして得る。そして、顔モジュール 30 は、各顔毎に、顔 ID（名前）34 及び顔方向 35 から、顔イベント生成部 36 により顔イベント 39 を生成して、ネットワークを介してアソシエーションモジュール 50 に対して送信するようになっている。

【0089】

上記顔ステレオモジュール 37 は、デバイス層としてのカメラ 15 と、プロセス層としての視差画像生成部 37a、目標抽出部 37b と、特徴層（データ）としての目標方向 37c と、イベント層としてのステレオイベント生成部 37d とから構成されている。これにより、ステレオモジュール 37 は、カメラ 15 からの画像信号に基づいて視差画像生成部 37a によって双方のカメラ 15 の画像信号から視差画像を生成する。次いで、目標抽出部 37b が、視差画像を領域分割し、その結果、縦に長い物体が発見されれば、目標抽出部 37b はそれを人物候補として抽出し、その目標方向 37c を決定（定位）する。ステレオイベント生成部 37d は、目標方向 37c に基づいてステレオイベント 39a を生成し、ネットワークを介してアソシエーションモジュール 50 に対して送信するようになっている。

【0090】

上記モータ制御モジュール 40 は、デバイス層としてのモータ 41 及びポテンシオメータ 42 と、プロセス層としての PWM 制御回路 43、AD 変換回路 44 及びモータ制御部 45 と、データである特徴層としてのロボット方向 46 と、イベント層としてのモタイイベント生成部 47 とから構成されている。これにより、モータ制御モジュール 40 においては、モータ制御部 45 がアテンション制御モジュール 57（後述）からの指令に基づいて PWM 制御回路 43 を介してモータ

タ 4 1 を駆動制御する。また、モータ 4 1 の回転位置をポテンシオメータ 4 2 により検出する。この検出結果は、A D 変換回路 4 4 を介してモータ制御部 4 5 に送られる。そして、モータ制御部 4 5 は、A D 変換回路 4 4 から受け取った信号からロボット方向 4 6 を抽出する。モータイベント生成部 4 7 は、ロボット方向 4 6 に基づいて、モータ方向情報から成るモータイベント 4 8 を生成して、ネットワークを介してアソシエーションモジュール 5 0 に対して送信するようになっている。

【0091】

上記アソシエーションモジュール 5 0 は、上述した聴覚モジュール 2 0、顔モジュール 3 0、ステレオモジュール 3 7、モータ制御モジュール 4 0 に対して、階層的に上位に位置付けられており、各モジュール 2 0、3 0、3 7、4 0 のイベント層の上位であるストリーム層を構成している。具体的には、アソシエーションモジュール 5 0 は、聴覚モジュール 2 0、顔モジュール 3 0、ステレオモジュール 3 7 及びモータ制御モジュール 4 0 からの非同期イベント 5 1、即ち聴覚イベント 2 9、顔イベント 3 9、ステレオイベント 3 9 a 及びモータイベント 4 8 を同期させて聴覚ストリーム 5 3、顔ストリーム 5 4、ステレオ視覚ストリーム 5 5 を生成する絶対座標変換部 5 2 と、各ストリーム 5 3、5 4、5 5 を関連付けてアソシエーションストリーム 5 9 を生成し、あるいはこれらストリーム 5 3、5 4、5 5 の関連付けを解除する関連付け部 5 6 と、さらにアテンション制御モジュール 5 7 と、ビューア 5 8 とを備えている。

【0092】

上記絶対座標変換部 5 2 は、聴覚モジュール 2 0 からの聴覚イベント 2 9、顔モジュール 3 0 からの顔イベント 3 9、ステレオモジュール 3 7 からのステレオイベント 3 9 a に、モータ制御モジュール 4 0 からのモータイベント 4 8 を同期させると共に、聴覚イベント 2 9、顔イベント 3 9 及びステレオイベント 3 9 a に関して、同期させたモータイベントによって、その座標系を絶対座標系に変換することにより、聴覚ストリーム 5 3、顔ストリーム 5 4 及びステレオ視覚ストリーム 5 5 を生成する。その際、上記絶対座標変換部 5 2 は、同一話者の聴覚ストリーム、顔ストリーム及びステレオ視覚ストリームに接続することによって、

聴覚ストリーム 53, 顔ストリーム 54 及びステレオ視覚ストリーム 55 を生成する。

【0093】

また、関連付け部 56 は、聴覚ストリーム 53, 顔ストリーム 54, ステレオ視覚ストリーム 55 に基づいて、これらのストリーム 53, 54, 55 の時間的つながりを考慮してストリームを関連付け、あるいは関連付けを解除して、アソシエーションストリーム 59 を生成すると共に、逆にアソシエーションストリーム 59 を構成する聴覚ストリーム 53, 顔ストリーム 54 及びステレオ視覚ストリーム 55 の結び付きが弱くなれば、関係付けを解除するようになっている。これにより、目標となる話者が移動している場合であっても、当該話者の移動を予測して、その移動範囲となる角度範囲内であれば、上述したストリーム 53, 54, 55 の生成を行なうことによって、当該話者の移動を予測して追跡できることになる。

【0094】

また、アテンション制御モジュール 57 は、モータ制御モジュール 40 の駆動モータ制御のプランニングのためのアテンション制御を行なうものであり、その際アソシエーションストリーム 59, 聴覚ストリーム 53, 顔ストリーム 54 そしてステレオ視覚ストリーム 55 の順に優先的に参照して、アテンション制御を行なう。そして、アテンション制御モジュール 57 は、聴覚ストリーム 53, 顔ストリーム 54 及びステレオ視覚ストリーム 55 の状態とアソシエーションストリーム 59 の存否に基づいて、ロボット 10 の動作プランニングを行ない、駆動モータ 41 の動作の必要があれば、モータ制御モジュール 40 に対して動作指令としてのモータイベントをネットワークを介して送信する。ここで、アテンション制御モジュール 57 におけるアテンション制御は、連続性とトリガに基づいており、連続性により同じ状態を保持しようとし、トリガにより最も興味のある対象を追跡しようとして、アテンションを向けるべきストリームを選択して、トラッキングを行なう。

【0095】

このようにして、アテンション制御モジュール 57 は、アテンション制御を行

なって、モータ制御モジュール 40 の駆動モータ 41 の制御のプランニングを行ない、このプランニングに基づいて、モータコマンド 64 a を生成し、ネットワーク 70 を介してモータ制御モジュール 40 に伝送する。これにより、モータ制御モジュール 40 では、このモータコマンド 64 a に基づいて、モータ制御部 45 が PWM 制御を行なって、駆動モータ 41 を回転駆動させて、ロボット 10 を所定方向に向けるようになっている。

【0096】

ビューア 58 は、このようにして生成された各ストリーム 53, 54, 55, 57 をサーバの画面上に表示するものであり、具体的にはレーダチャート 58 a 及びストリームチャート 58 b により表示する。レーダチャート 58 a は、その瞬間におけるストリームの状態、より詳細にはカメラの視野角と音源方向を示し、ストリームチャート 58 b は、アソシエーションストリーム（太線図示）と聴覚ストリーム、顔ストリーム及びステレオ視覚ストリーム（細線図示）を示している。

【0097】

本発明実施形態による人型ロボット 10 は以上のように構成されており、以下のように動作する。

まず、ロボット 10 の前方 1 m の距離で、斜め左 ($\theta = +60$ 度), 正面 ($\theta = 0$ 度) そして斜め右 ($\theta = -60$ 度) の方向に、それぞれ話者が並んでおり、ロボット 10 が対話部 28 により、三人の話者に質問して、各話者が同時に質問に対する回答を行なう。

【0098】

これにより、ロボット 10 はマイク 16 が当該話者の音声を拾って、聴覚モジュール 20 が音源方向を伴う聴覚イベント 29 を生成して、ネットワークを介してアソシエーションモジュール 50 に伝送する。これにより、アソシエーションモジュール 50 は、この聴覚イベント 29 に基づいて、聴覚ストリーム 53 を生成する。

【0099】

また、顔モジュール 30 は、カメラ 15 による話者の顔の画像を取り込んで、

顔イベント 39 を生成して、当該話者の顔を顔データベース 38 により検索し、顔識別を行なうと共に、その結果である顔 ID 24 及び画像をネットワーク 70 を介してアソシエーションモジュール 50 に伝送する。なお、当該話者の顔が顔データベース 38 に登録されていない場合には、顔モジュール 30 は、その旨をネットワークを介してアソシエーションモジュール 50 に伝送する。

【0100】

従って、アソシエーションモジュール 50 は、これらの聴覚イベント 29、顔イベント 39、ステレオイベント 39a に基づいて、アソシエーションストリーム 59 を生成する。

【0101】

ここで、聴覚モジュール 20 は、アクティブ方向通過型フィルタ 23a により、聴覚エピソード幾何による IPD を利用して、各音源（話者 X, Y, Z）の定位及び分離を行なって、分離音（音響信号）を取り出す。そして、聴覚モジュール 20 は、その音声認識部 27 により音声認識エンジン 27c を使用して、各話者 X, Y, Z の音声を認識してその結果を対話部 28 に出力する。これにより、対話部 28 は、音声認識部 27 により音声認識された前記回答を、それぞれの話者に対してロボット 10 が正対した状態で発話する。なお、音声認識部 27 が正しく音声認識できなかった場合には、ロボット 10 が当該話者に正対した状態で再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【0102】

このようにして、本発明実施形態による人型ロボット 10 によれば、聴覚モジュール 20 により音源定位・音源分離された分離音（音響信号）に基づいて、音声認識部 27 が、各話者及び方向に対応する音響モデルを使用して音声認識を行なうことにより同時に発話する複数の話者の音声を音声認識することができる。

【0103】

以下に、音声認識部 27 の動作を実験により評価する。

これらの実験においては、図 8 に示すように、ロボット 10 の前方 1 m の距離で、斜め左（ $\theta = +60$ 度）、正面（ $\theta = 0$ 度）そして斜め右（ $\theta = -60$ 度）の方向に、それぞれ話者 X, Y, Z が並んでいる。なお、実験では、話者として

人間の代わりにそれぞれスピーカを置くと共に、その前面に話者の写真を配置している。ここで、スピーカは、音響モデルを作成したときと同じスピーカを使用しており、スピーカから発せられた音声を写真の話者の音声とみなしている。

【0104】

そして、以下のシナリオに基づいて音声認識の実験を行なう。

1. ロボット10が三人の話者X, Y, Zに質問する。
2. 三人の話者X, Y, Zが同時に質問に対する回答を行なう。
3. ロボット10が三人の話者X, Y, Zの混合音声に基づいて、音源定位・音源分離を行ない、さらに各分離音について音声認識を行なう。
4. ロボット10が、順次に各話者X, Y, Zに正対した状態で当該話者の回答を答える。
5. ロボット10は、音声認識が正しくできなかったと判断したとき、当該話者に正対して再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【0105】

上記シナリオによる実験結果の第一の例を図9に示す。

1. ロボット10が「好きな数字は何ですか?」と質問する。(図9(a)参照)
2. 各話者X, Y, Zとしてのスピーカから、同時に1から10までの数字のうちから、任意の数字を読み上げた音声を流す。例えば図9(b)に示すように、話者Xは「2」、話者Yは「1」そして話者Zは「3」と言う。
3. ロボット10は、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって、分離音を抽出する。そして、各話者X, Y, Zに対応する分離音に基づいて、各話者別に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。
4. その際、音声認識部27のセクタ27eが、正面が話者Yであると仮定して音声認識の評価を行ない(図9(c))、続いて正面が話者Xであると仮定して音声認識の評価を行ない(図9(d))、最後に正面が話者Zであると仮定して音声認識の評価を行なう(図9(e))。

5. そして、セクタ 27 e が、音声認識結果を統合して、図 9 (f) に示すように、ロボット正面 ($\theta = 0$ 度) に関して、最も適合の良い話者名 (Y) と音声認識結果 (「1」) を決定し対話部 28 に出力する。これにより、図 9 (g) に示すように、ロボット 10 が話者 Y に正対した状態にて、「Y さんは「1」です。」と答える。

6. 続いて、斜め左 ($\theta = +60$ 度) の方向に関して、上記と同様の処理を行って、図 9 (h) に示すように、ロボット 10 が話者 X に正対した状態にて、「X さんは「2」です。」と答える。更に、斜め右 ($\theta = -60$ 度) の方向に対しても同様の処理を行って、図 9 (i) に示すように、ロボット 10 が話者 Z に正対した状態にて、「Z さんは「3」です。」と答える。

【0106】

この場合、ロボット 10 は、各話者 X, Y, Z の回答をすべて正しく音声認識することができた。従って、同時発話の場合であっても、ロボット 10 のマイク 16 を使用したロボット視聴覚システム 17 における音源定位・音源分離・音声認識の有効性が示された。

【0107】

なお、図 9 (j) に示すように、ロボット 10 が各話者に正対せずに、「Y さんは「1」です。X さんは「2」です。Z さんは「3」です。合計「6」です。」というように、各話者 X, Y, Z の答えた数字の合計も答えるようにしてもよい。

【0108】

図 10 は、上述したシナリオによる実験結果の第二の例を示している。

1. 図 9 に示した第一の例と同様に、ロボット 10 が「好きな数字は何ですか?」と質問し (図 10 (a) 参照)、各話者 X, Y, Z としてのスピーカから、図 10 (b) に示すように、話者 X は「2」、話者 Y は「1」そして話者 Z は「3」という音声の流れる。

2. ロボット 10 は、同様に、聴覚モジュール 20 にて、そのマイク 16 で集音した音響信号に基づいて、アクティブ方向通過型フィルタ 23 a により音源定位・音源分離を行なって分離音を抽出し、各話者 X, Y, Z に対応する分離

音に基づいて、各話者別に音声認識部 27 が 9 つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。その際、音声認識部 27 のセクタ 27 e は、図 10 (c) に示すように、正面の話者 Y については、正しく音声認識の評価を行なうことができる。

3. これに対して、+60 度に位置する話者 X について、セクタ 27 e は、図 10 (d) に示すように、「2」であるか「4」であるか決定することができない。

4. 従って、ロボット 10 は、図 10 (e) に示すように、+60 度に位置する話者 X に正対して、「2 ですか？ 4 ですか？」と質問する。

5. これに対して、図 10 (f) に示すように、話者 X であるスピーカから「2」という回答が流れる。この場合、話者 X は、ロボット 10 の正面に位置していることから、聴覚モジュール 20 が話者 X の回答について正しく音源定位・音源分離し、音声認識部 27 が正しく音声認識して、話者名 X と音声認識結果「2」を対話部 28 に出力する。これにより、ロボット 10 は、図 10 (g) に示すように、話者 X に対して「X さんは「2」です。」と答える。

6. 続いて、話者 Z についても同様の処理を行なって、その音声認識結果を話者 Z に対して答える。即ち、図 10 (h) に示すように、ロボット 10 が話者 Z に正対した状態にて、「Z さんは「3」です。」と答える。

【0109】

このようにして、ロボット 10 は、再質問により、各話者 X, Y, Z の回答をすべて正しく音声認識することができた。従って、側方での聴覚中心窩の影響による分離精度の低下による音声認識の曖昧さを、ロボット 10 が側方の話者に対して正対して再質問することにより解消して、音源分離精度を向上させ、音声認識精度を向上させることができることが示された。

【0110】

なお、図 10 (i) に示すように、ロボット 10 が各話者の音声認識を正しく行なった後、「Y さんは「1」です。X さんは「2」です。Z さんは「3」です。合計「6」です。」というように、各話者 X, Y, Z の答えた数字の合計も答えるようにしてもよい。

【0111】

図11は、上述したシナリオによる実験結果の第三の例を示している。

1. この場合も図9に示した第一の例と同様にして、ロボット10が「好きな数字は何ですか？」と質問し（図10（a）参照）、各話者X, Y, Zとしてのスピーカから、図10（b）に示すように、話者Xは「8」、話者Yは「7」そして話者Zは「9」という音声流れる。

2. ロボット10は、同様にして、聴覚モジュール20にて、そのマイク16で集音した音響信号に基づいて、リアルタイムトラッキング（X3'参照）によるストリーム方向 θ 、そして各話者の顔イベントを参照して、アクティブ方向通過型フィルタ23aにより音源定位・音源分離を行なって分離音を抽出し、各話者X, Y, Zに対応する分離音に基づいて、各話者毎に音声認識部27が9つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

その際、音声認識部27のセクタ27eは、正面の話者Yについては、顔イベントに基づいて話者Yである確率が高いことから、各音響モデルによる音声認識結果の統合の際に、図10（c）に示すようにこれを考慮する。これにより、より正確な音声認識を行なうことができる。従って、ロボット10は、図11（d）に示すように、話者Xに対して「Xさんは「7」です。」と答える。

3. これに対して、+60度に位置する話者Xについて、ロボット10が向きを変えて正対すると、このときの正面の話者Xについて、顔イベントに基づいて話者Xである確率が高いので、同様にして、セクタ27eは、図11（e）に示すようにこれを考慮する。従って、ロボット10は、図11（f）に示すように、話者Xに対して「Yさんは「8」です。」と答える。

4. 続いて、セクタ27eは、図11（g）に示すように、話者Zについても同様の処理を行なって、その音声認識結果を話者Zに対して答える。即ち、図11（h）に示すように、ロボット10が話者Zに正対した状態にて「Zさんは「9」です。」と答える。

【0112】

このようにして、ロボット10は、各話者毎に正対して、その顔イベントを参

照しながら話者の顔認識に基づいて、各話者 X, Y, Z の回答をすべて正しく音声認識することができた。これにより、顔認識により話者が誰であることを特定することができるので、より精度の高い音声認識を行なうことができることが示された。特に、特定の環境での利用を前提とするような場合、顔認識によってほぼ 100% に近い顔認識精度が得られると、顔認識情報を信頼性の高い情報として利用することができることになり、音声認識部 27 の音声認識エンジン 27c で使用される音響モデル 27d の数を削減することができるので、より高速で且つ高精度の音声認識が可能になる。

【0113】

図 12 は、上述したシナリオによる実験結果の第四の例を示している。

1. ロボット 10 が「好きなフルーツは何ですか？」と質問し（図 12 (a) 参照）、各話者 X, Y, Z としてのスピーカから、例えば図 12 (b) に示すように、話者 X は「梨」、話者 Y は「スイカ」そして話者 Z は「メロン」と言う。

2. ロボット 10 は、聴覚モジュール 20 にて、そのマイク 16 で集音した音響信号に基づいて、アクティブ方向通過型フィルタ 23a により音源定位・音源分離を行なって分離音を抽出する。そして、各話者 X, Y, Z に対応する分離音に基づいて、各話者毎に音声認識部 27 が 9 つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

3. その際、音声認識部 27 のセクタ 27e が、正面が話者 Y であると仮定して音声認識の評価を行ない（図 12 (c)）、続いて正面が話者 X であると仮定して音声認識の評価を行ない（図 12 (d)）、最後に正面が話者 Z であると仮定して音声認識の評価を行なう（図 12 (e)）。

4. そして、セクタ 27e が、音声認識結果を統合して、図 12 (f) に示すように、ロボット正面 ($\theta = 0$ 度) 方向に関して最も適合の良い話者名 (Y) と音声認識結果 (「スイカ」) を決定し対話部 28 に出力する。これにより、図 9 (g) に示すように、ロボット 10 が話者 Y に正対した状態にて、「Y さんは「スイカ」です。」と答える。

5. 続いて、各話者 X, Z についても同様の処理を行なって、その音声認識結果を各話者 X, Z に対して答える。即ち、図 12 (h) に示すように、ロボット

10が話者Xに正対した状態にて、「Xさんは「梨」です。」と答え、さらに図12(i)に示すように、ロボット10が話者Zに正対した状態にて「Zさんは「メロン」です。」と答える。

【0114】

この場合、ロボット10は、各話者X、Y、Zの回答をすべて正しく音声認識することができた。従って、音声認識エンジン27cに登録された単語は数字に限ることなく、前もって登録された単語であれば、音声認識可能であることが分かる。ここで、実験に使用した音声認識エンジン27cでは、約150語の単語が登録されている。なお、単語の音節数が多くなると、音声認識率はやや低くなる。

【0115】

上述した実施形態においては、ロボット10は、その上半身が4DOF（自由度）を有するように構成されているが、これに限らず、任意の動作を行なうように構成されたロボットに本発明によるロボット視聴覚システムを組み込むことも可能である。

【0116】

また、上述した実施形態においては、本発明によるロボット視聴覚システムを人型ロボット10に組み込んだ場合について説明したが、これに限らず、犬型等の各種動物型ロボットや、その他の形式のロボットに組み込むことも可能であることは明らかである。

【0117】

また、上記説明では、図4に示すようにロボット視聴覚システム17がステレオモジュール37を備える構成例を説明したが、本発明の実施形態に係るロボット視聴覚システムは、ステレオモジュール37を備えずに構成することもできる。この場合、アソシエーションモジュール50は、聴覚イベント29、顔イベント39及びモータイベント48に基づいて、各話者の聴覚ストリーム53及び顔ストリーム54を生成し、さらにこれらの聴覚ストリーム53及び顔ストリーム54を関連付けてアソシエーションストリーム59を生成するように構成され、アテンション制御モジュール50においては、これらのストリームに基づいてア

テンション制御が行われるように構成される。

【0118】

さらに、上記説明においては、アクティブ方向通過型フィルタ 23a は、方向毎に通過帯域幅（パスレンジ）を制御しており、処理する音の周波数によらず通過帯域幅を一定としていた。

ここで、通過帯域 δ を導出するために、100Hz, 200Hz, 500Hz, 1000Hz, 2000Hz, 100Hz の調波構造音（ハーモニクス）の 5 つの純音と 1 つのハーモニクスとを用いて、1 音源に対する音源抽出率を調べる実験を行った。なお、音源をロボット正面である 0 度からロボットの左位置或いは右位置である 90 度の範囲で 10 度毎に位置を移動させた。図 13～図 15 は音源を 0 度から 90 度の範囲の各位置に設置した場合の音源抽出率を示すグラフであり、この実験結果が示すように、周波数に応じて通過帯域幅を制御することにより、特定の周波数の音の抽出率を向上させることができ、分離精度を向上できる。よって、音声認識率も向上する。従って、上記説明したロボット視聴覚システム 17 においては、アクティブ方向通過型フィルタ 23a のパスレンジが、周波数毎に制御可能に構成されるのが望ましい。

【0119】

【発明の効果】

以上述べたように、この発明によれば、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能である。また、各音響モデルによる音声認識結果をセクタにより統合して、最も信頼性の高い音声認識結果を判断することにより、従来の音声認識に比べて、より一層正確な音声認識を行なうことができる。

【図面の簡単な説明】

【図 1】

この発明によるロボット聴覚装置の第一の実施形態を組み込んだ人型ロボットの外観を示す正面図である。

【図 2】

図 1 の人型ロボットの側面図である。

【図 3】

図 1 の人型ロボットにおける頭部の構成を示す概略拡大図である。

【図 4】

図 1 の人型ロボットにおけるロボット視聴覚システムの電氣的構成例を示すブロック図である。

【図 5】

図 4 に示すロボット視聴覚システムにおける聴覚モジュールの作用を示す図である。

【図 6】

図 4 のロボット視聴覚システムにおける聴覚モジュールの音声認識部で 사용되는音声認識エンジンの構成例を示す概略斜視図である。

【図 7】

図 6 の音声認識エンジンによる正面及び左右±60度の方向の話者による音声の認識率を示すグラフであり、(A)は正面の話者、(B)は斜め左+60度の話者そして(C)は斜め右-60度の話者の場合を示している。

【図 8】

図 4 に示すロボット視聴覚システムにおける音声認識実験を示す概略斜視図である。

【図 9】

図 4 のロボット視聴覚システムの音声認識実験の第一の例の結果を順次を示す図である。

【図 10】

図 4 のロボット視聴覚システムの音声認識実験の第二の例の結果を順次を示す図である。

【図 11】

図 4 のロボット視聴覚システムの音声認識実験の第三の例の結果を順次を示す図である。

【図 12】

図 4 のロボット視聴覚システムの音声認識実験の第四の例の結果を順次を示す

図である。

【図 1 3】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は0度、(b)は10度、(c)は20度、(d)は30度の方向に音源がある場合である。

【図 1 4】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は40度、(b)は50度、(c)は60度の方向に音源がある場合である。

【図 1 5】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、(a)は70度、(b)は80度、(c)は90度の方向に音源がある場合である。

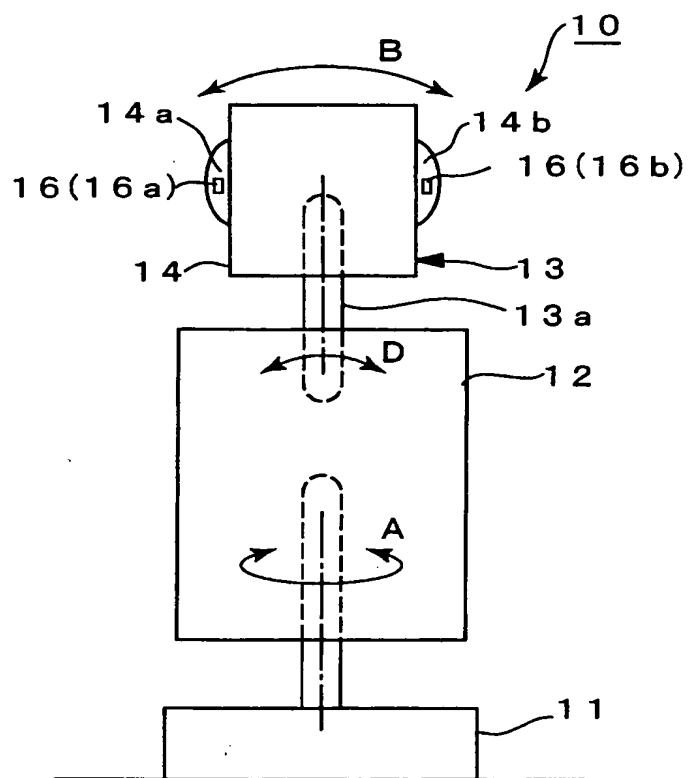
【符号の説明】

- 1 0 人型ロボット
- 1 1 ベース
- 1 2 胴体部
- 1 3 頭部
- 1 4 外装
- 1 5 カメラ (ロボット視覚)
- 1 6, 1 6 a, 1 6 b マイク (ロボット聴覚)
- 1 7 ロボット視聴覚システム
- 2 0 聴覚モジュール
- 2 1 ピーク抽出部
- 2 2 音源定位部
- 2 3 音源分離部
- 2 3 a アクティブ方向通過型フィルタ
- 2 6 聴覚イベント生成部
- 2 7 音声認識部

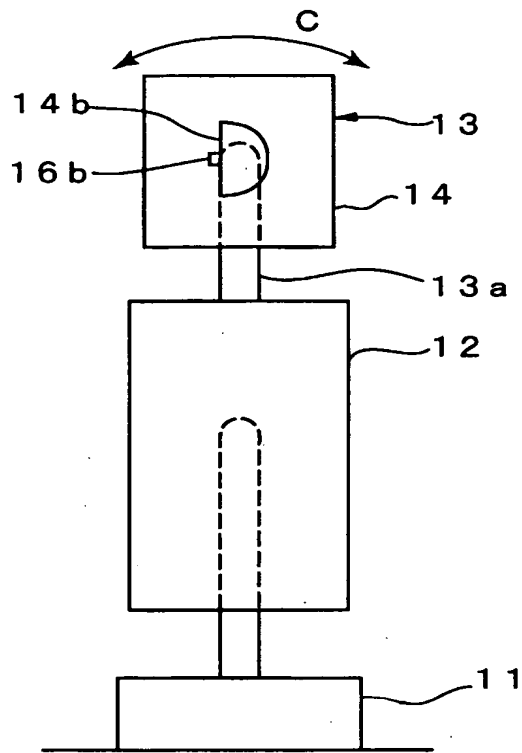
- 2 7 a 自声抑制部
- 2 7 b 自動認識部
- 2 7 c 音声認識エンジン
- 2 7 d 音響モデル
- 2 7 e セレクタ
- 2 8 対話部
- 3 0 顔モジュール
- 3 7 ステレオモジュール
- 4 0 モータ制御モジュール
- 5 0 アソシエーションモジュール
- 5 7 アテンション制御モジュール

【書類名】 図面

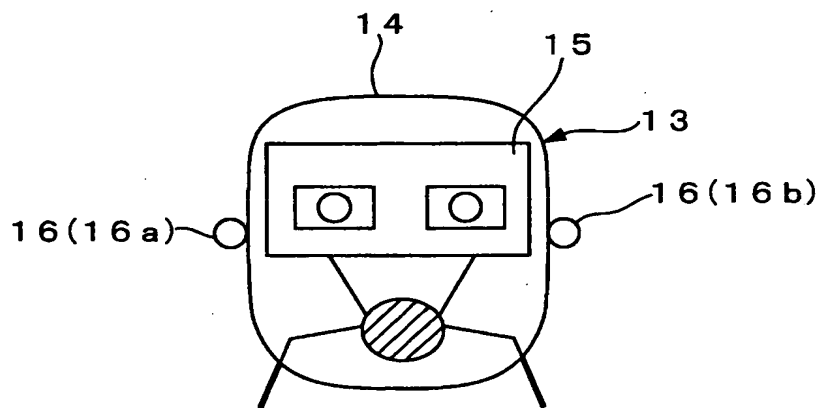
【図 1】



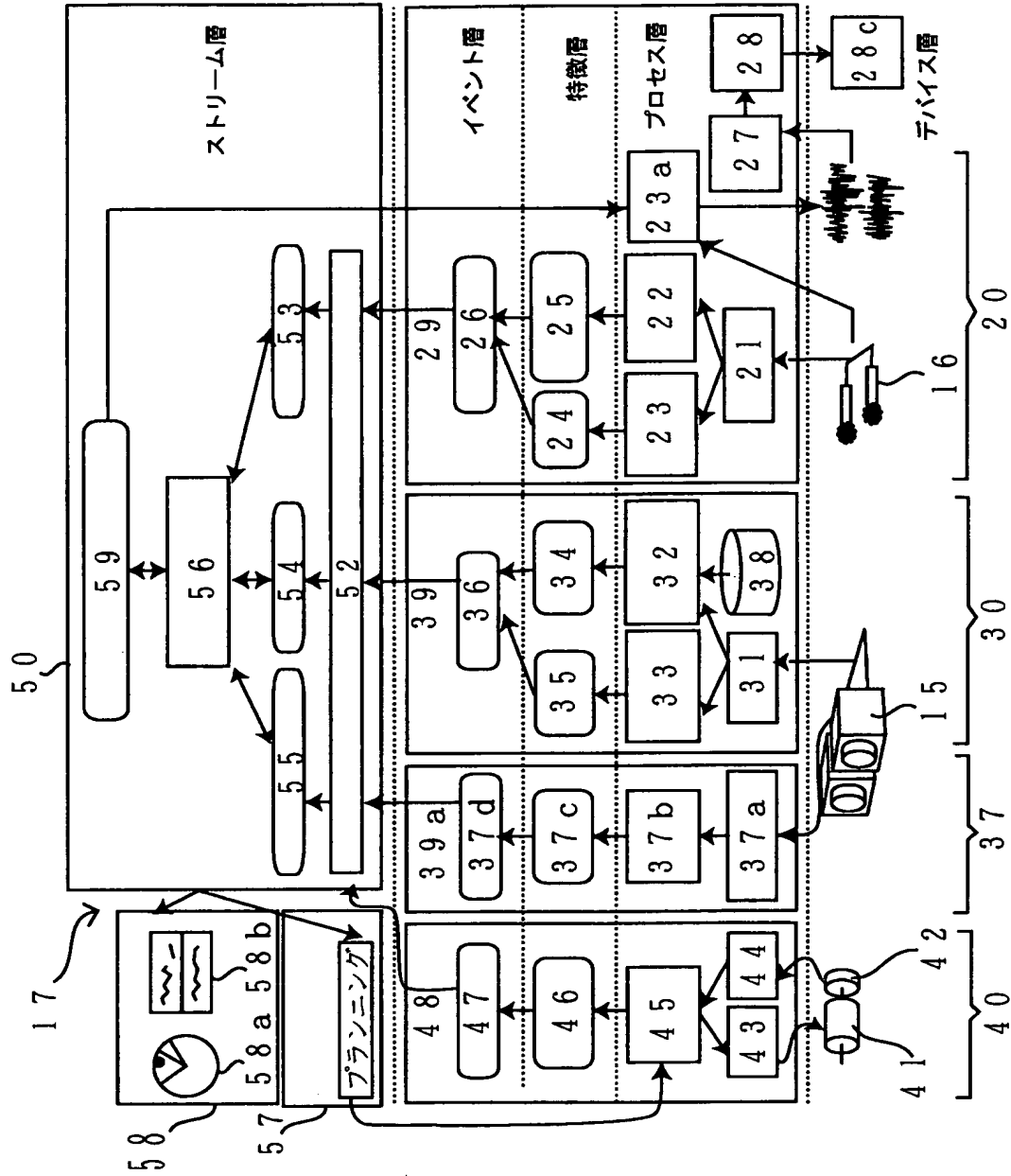
【図 2】



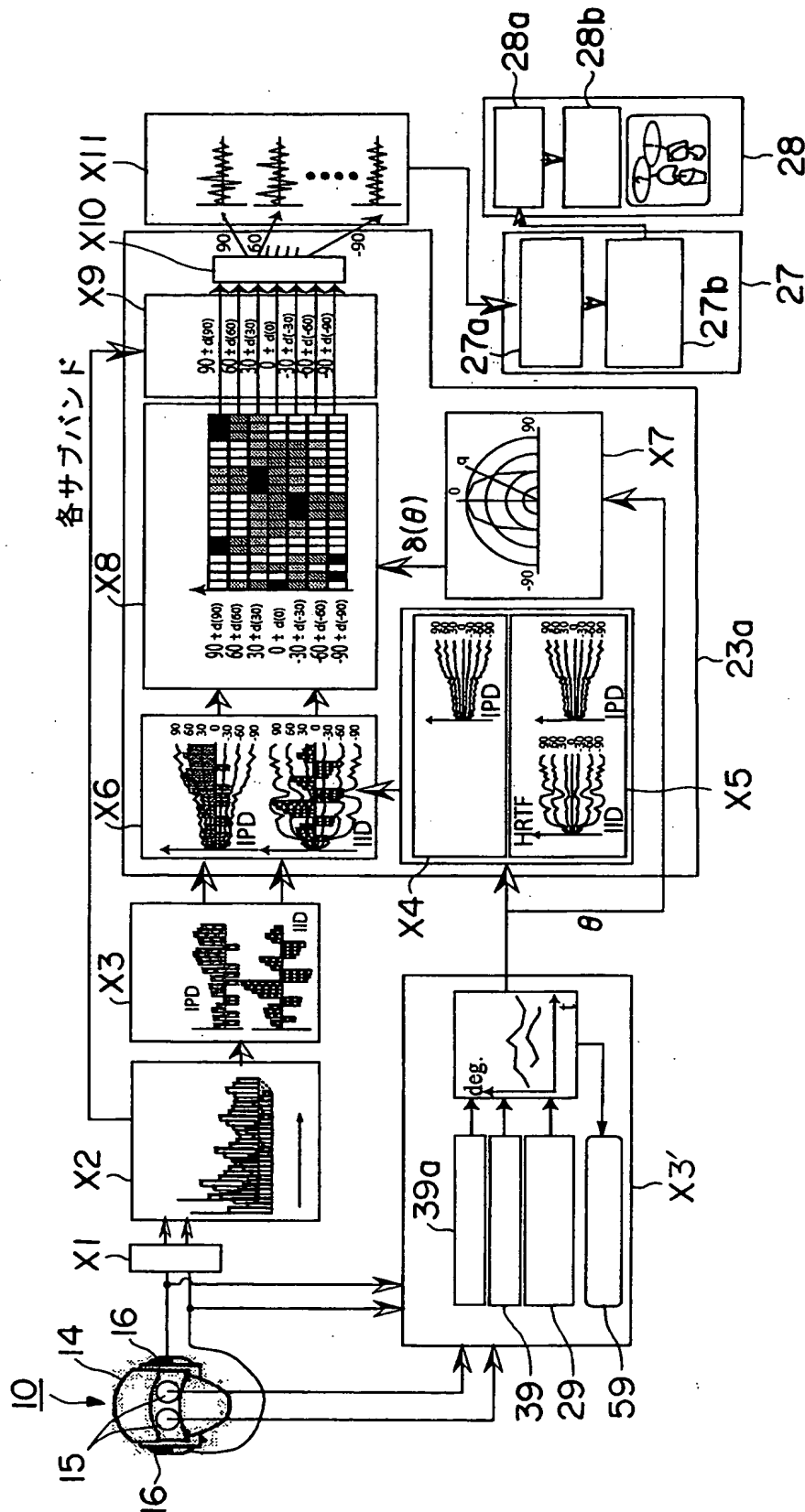
【図 3】



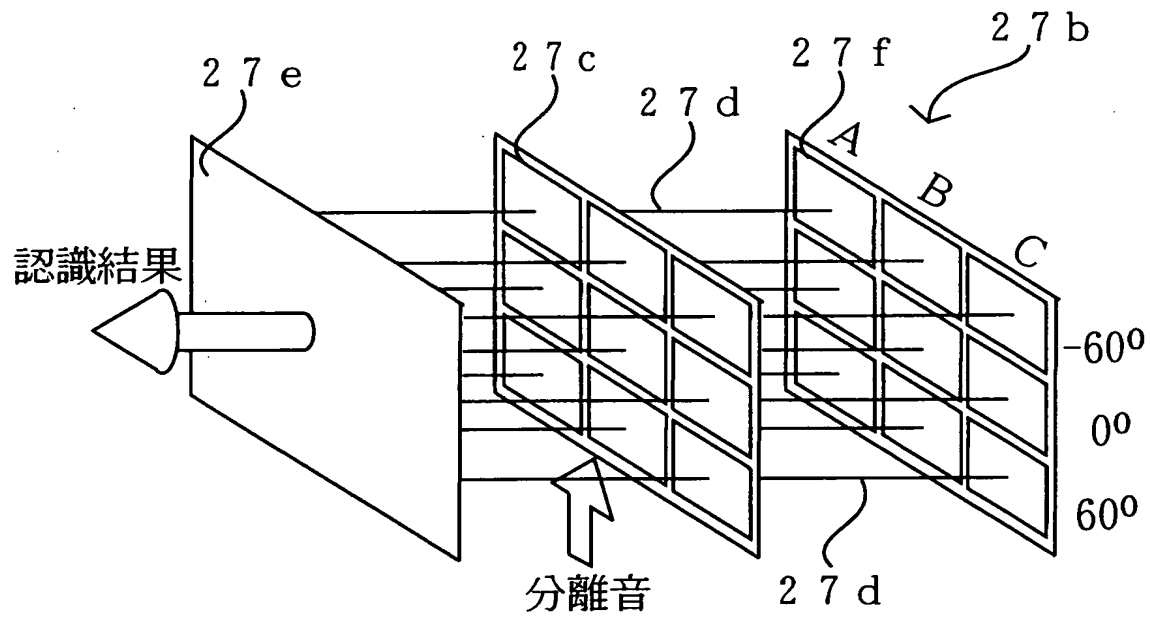
【図 4】



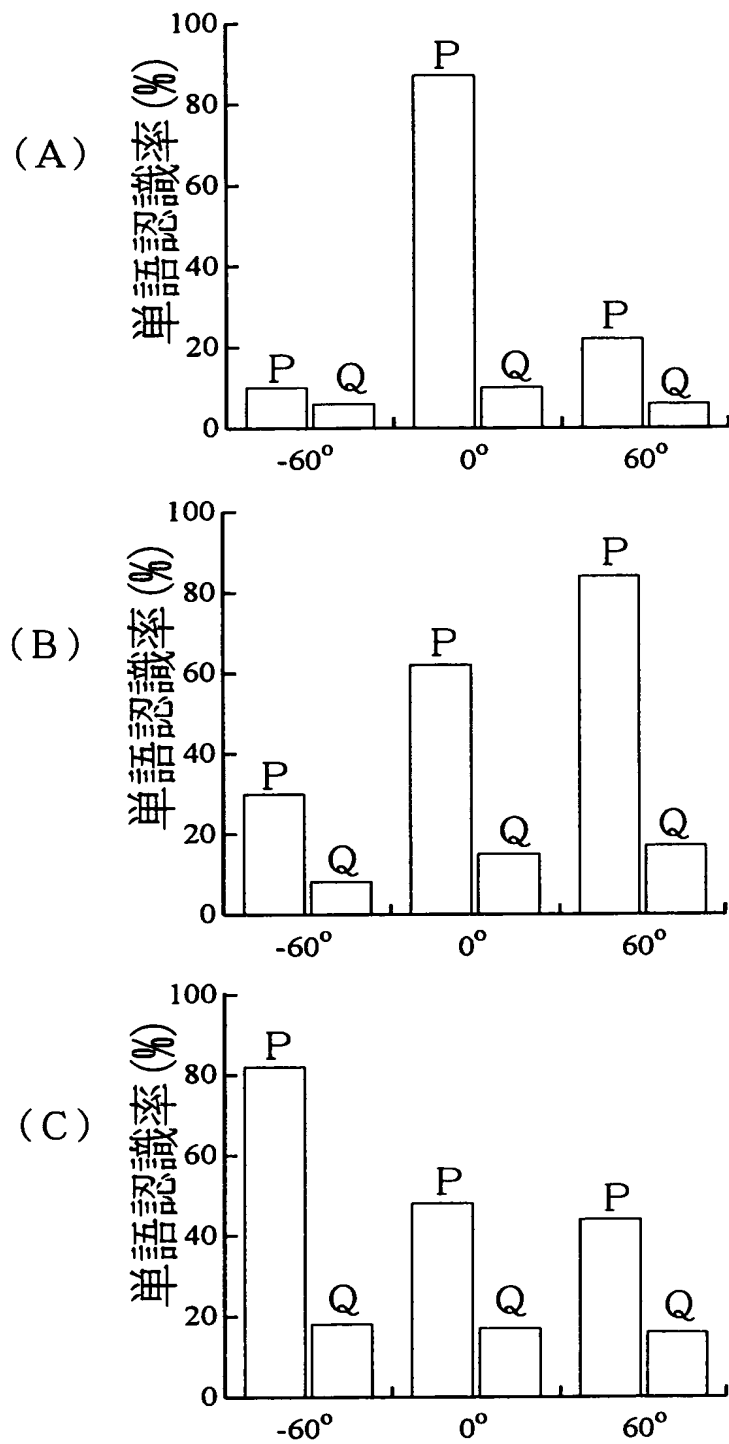
【図 5】



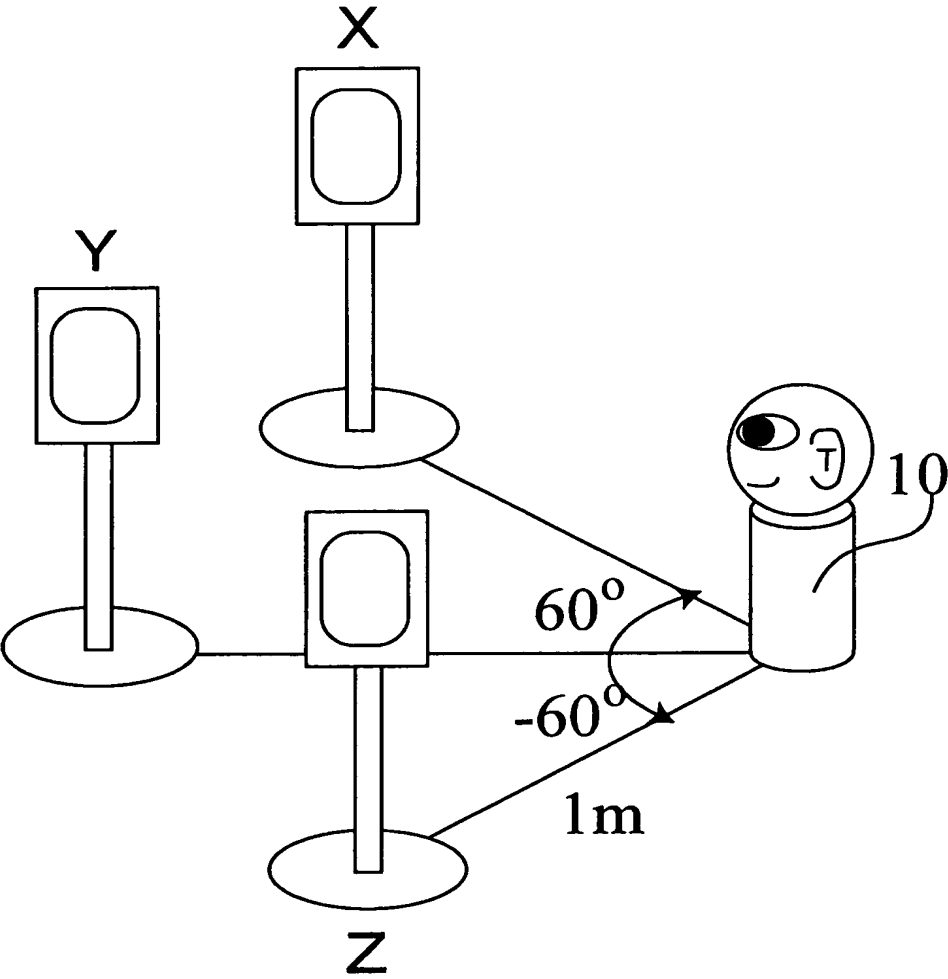
【図 6】



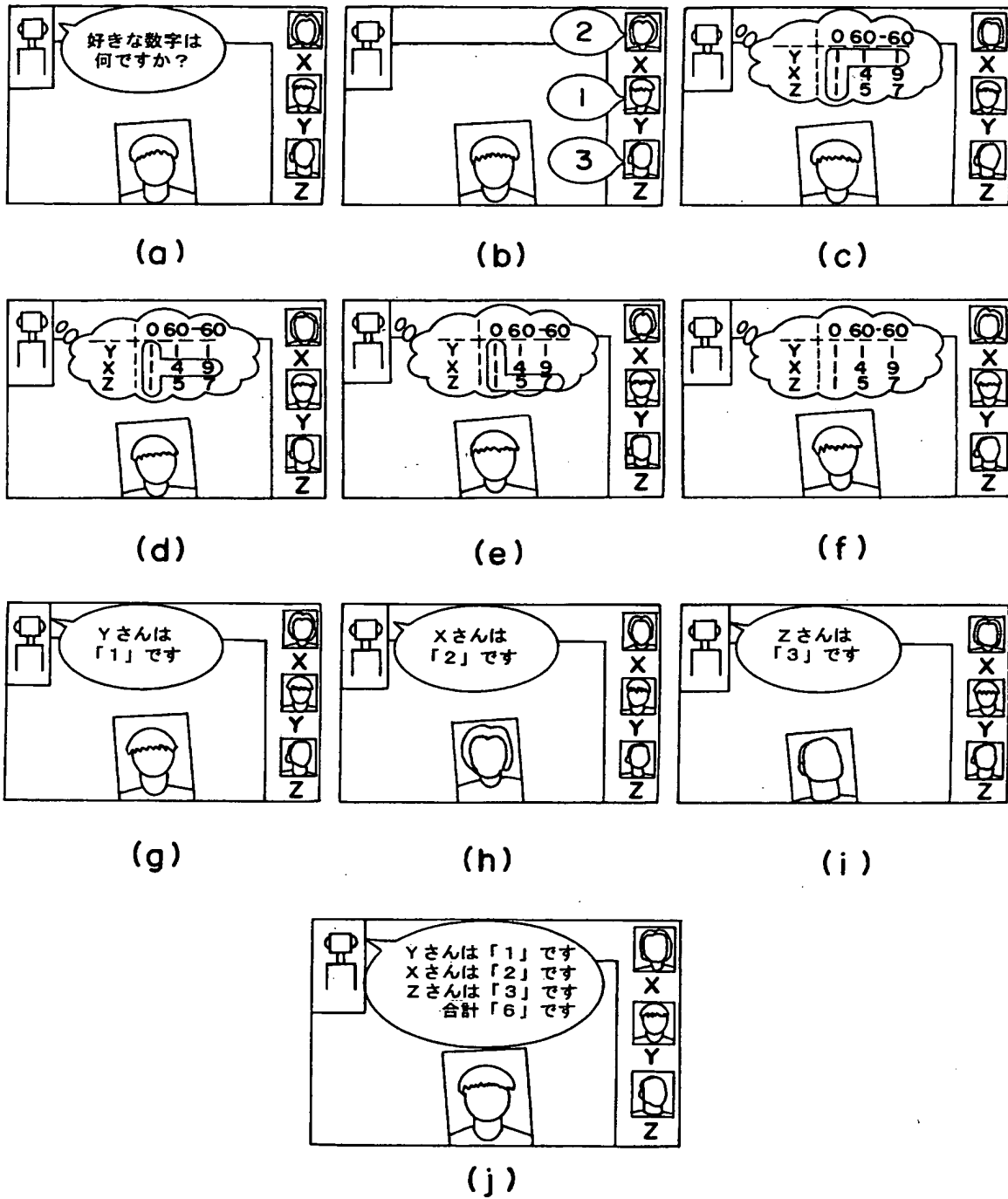
【図 7】



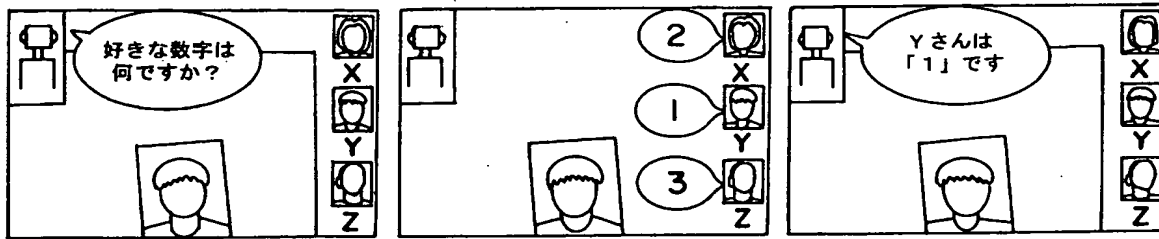
【図 8】



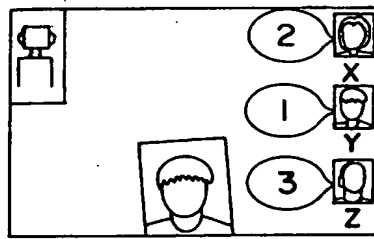
【図 9】



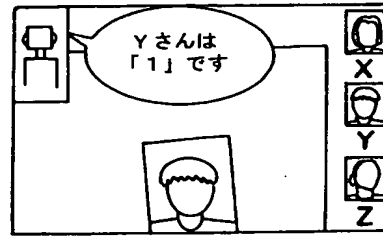
【図 10】



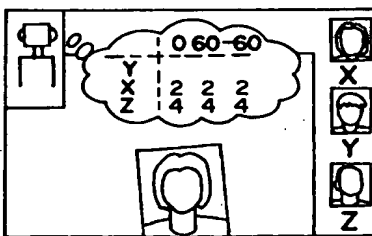
(a)



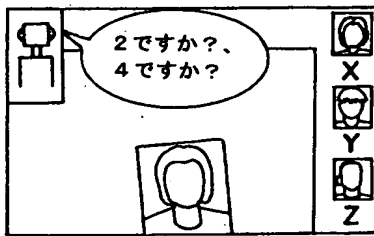
(b)



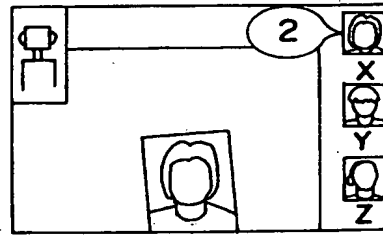
(c)



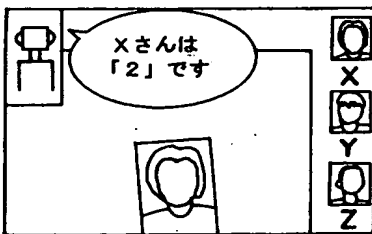
(d)



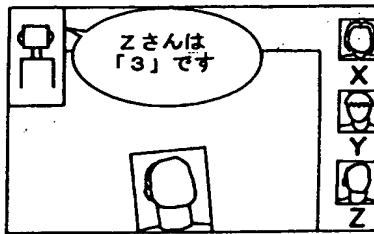
(e)



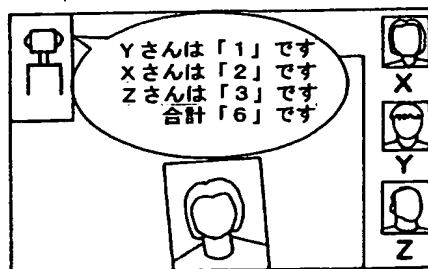
(f)



(g)

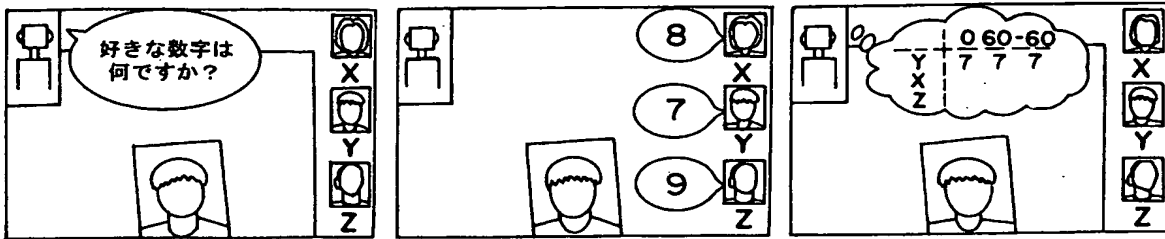


(h)



(i)

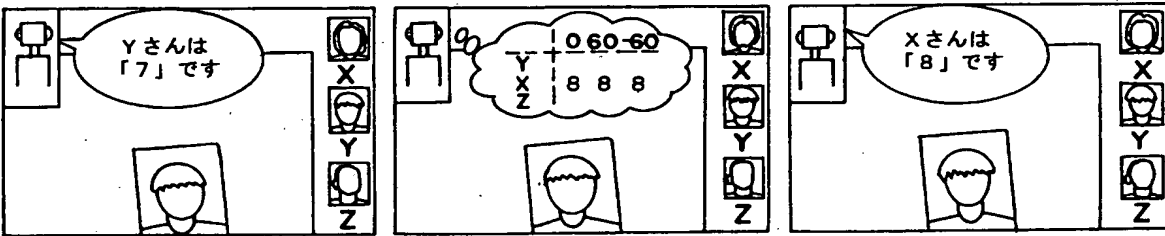
【図 11】



(a)

(b)

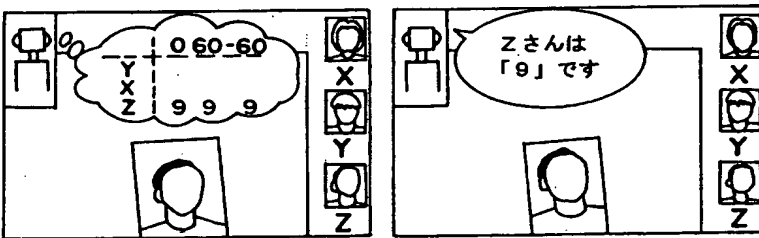
(c)



(d)

(e)

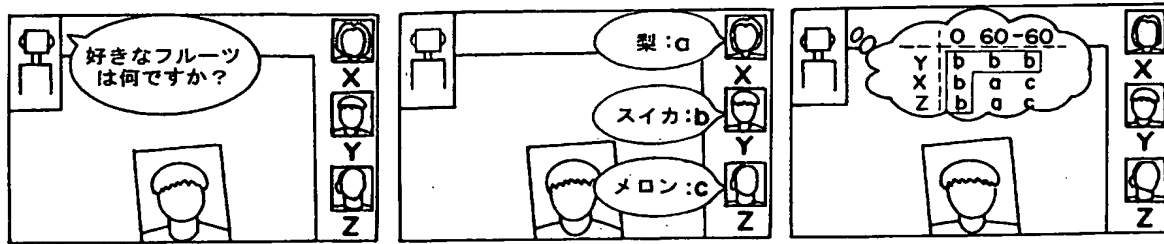
(f)



(g)

(h)

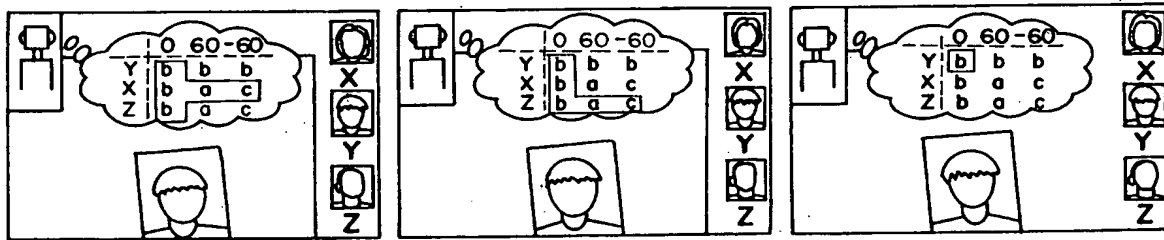
【図 12】



(a)

(b)

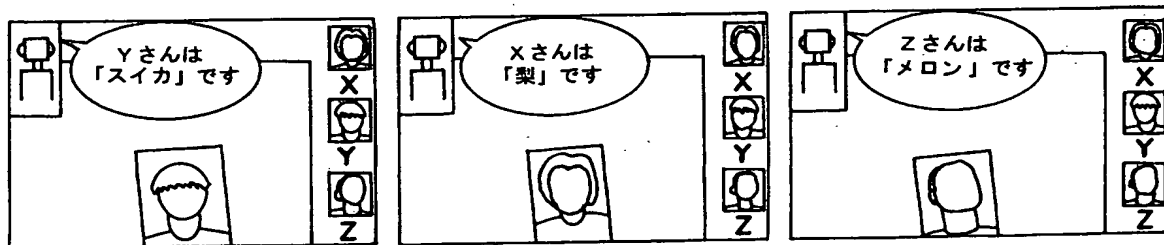
(c)



(d)

(e)

(f)

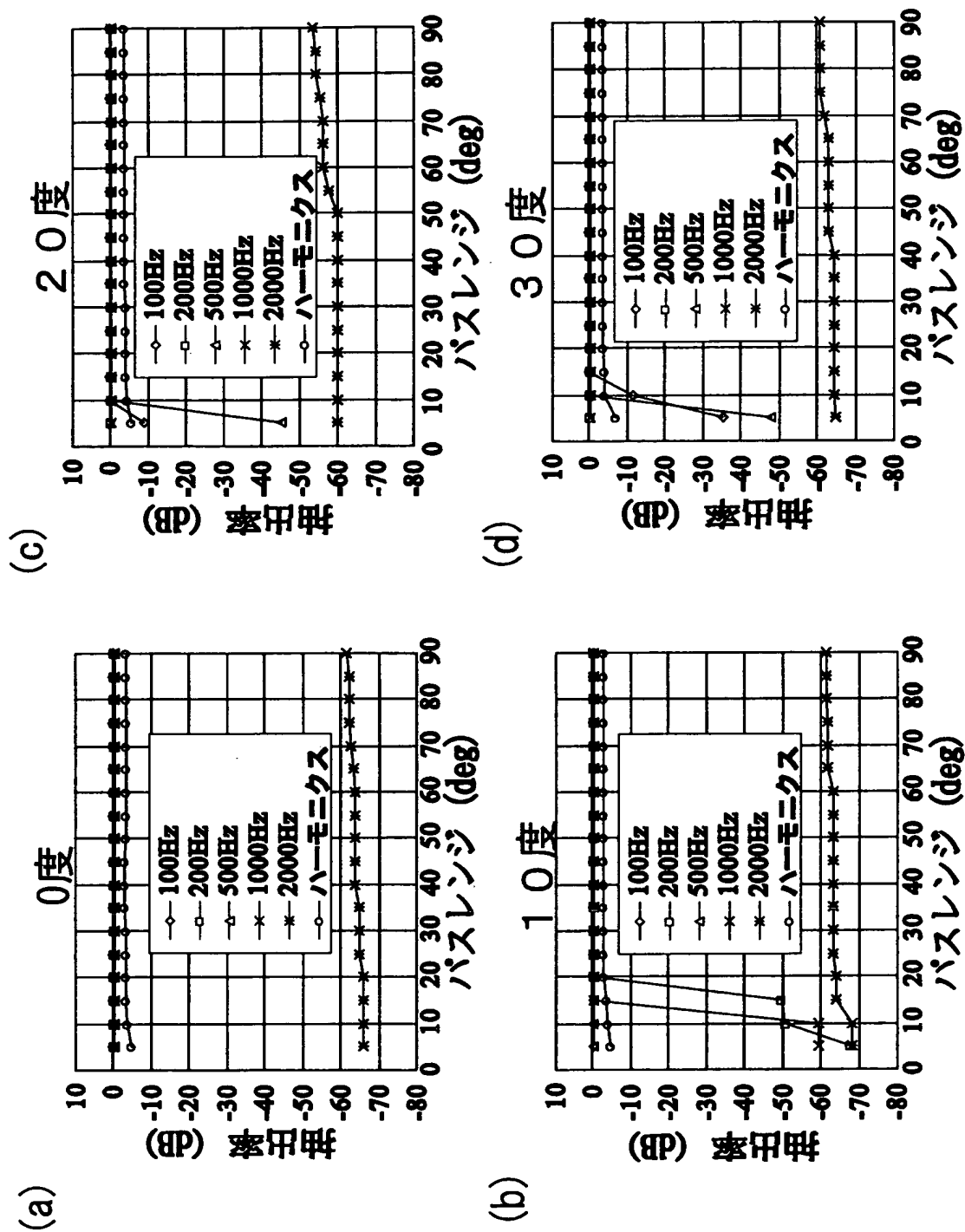


(g)

(h)

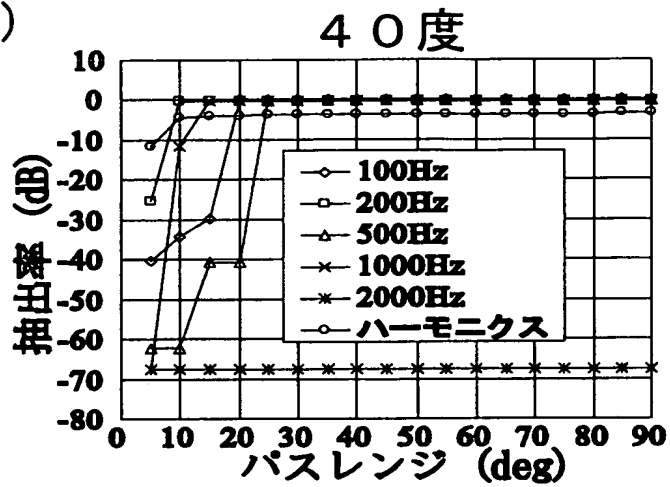
(i)

【図 13】

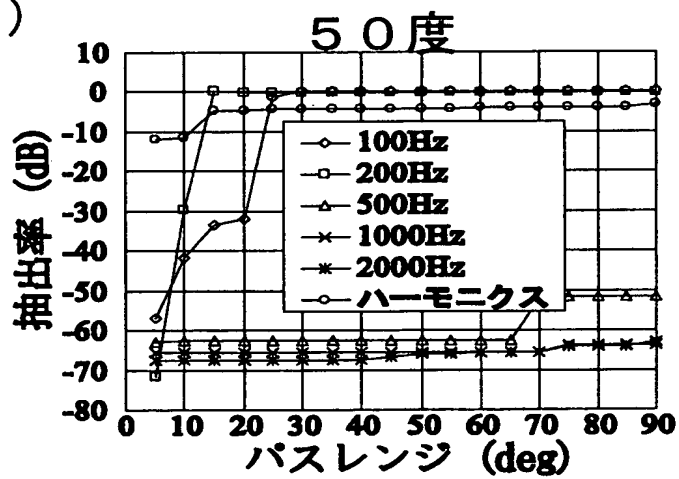


【図 14】

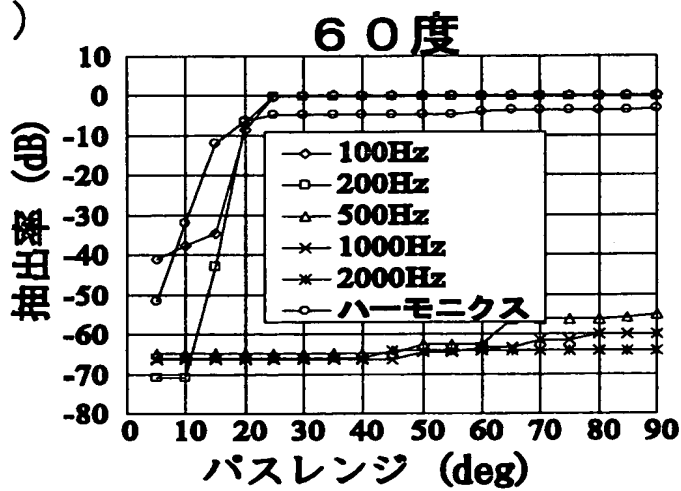
(a)



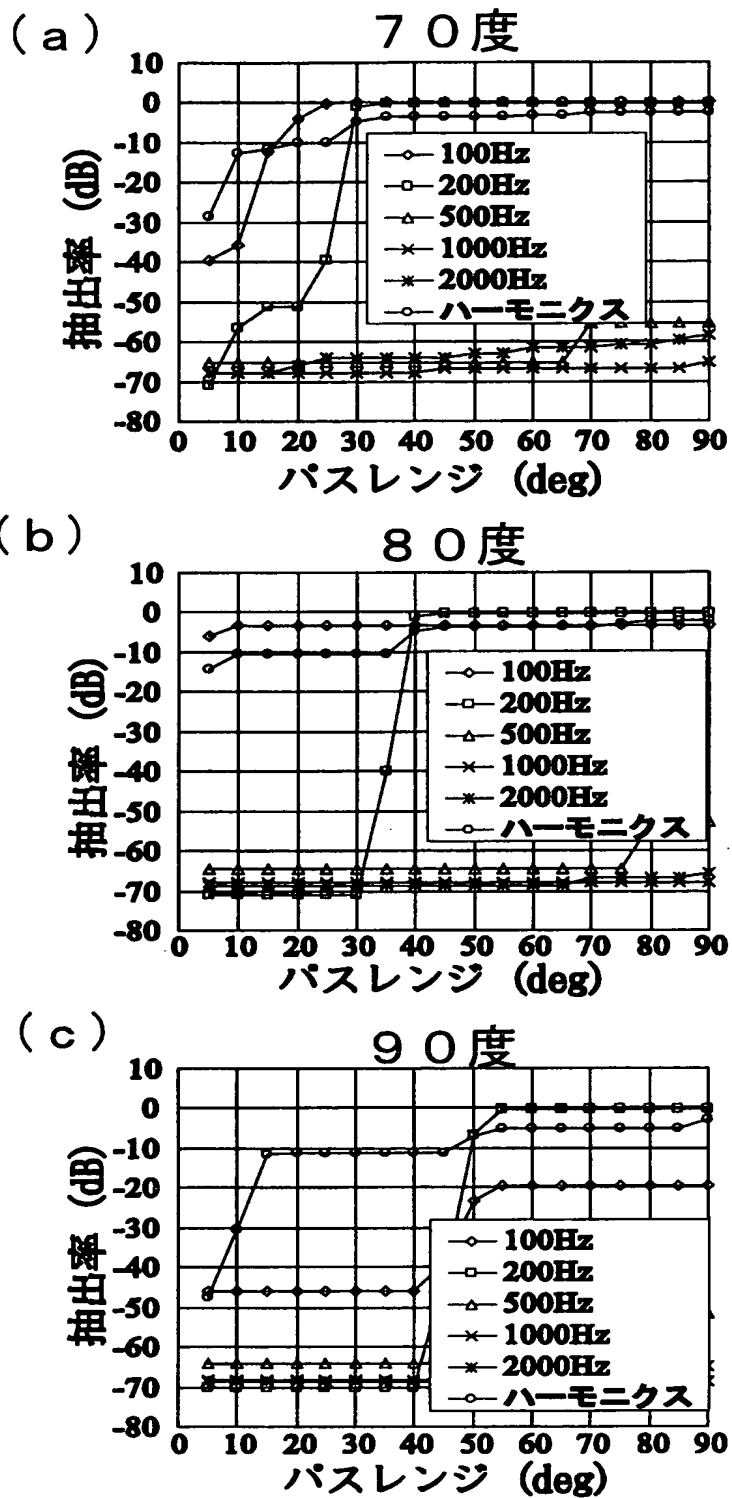
(b)



(c)



【図15】



【書類名】 要約書

【要約】

【課題】 各音源からの分離された音についての認識を行なうようにしたロボット視聴覚システムを提供する。

【解決手段】 聴覚モジュール 2 0, 顔モジュール 3 0, ステレオモジュール 3 7, モータ制御モジュール 4 0 と、これらの各モジュールを制御するアソシエーションモジュール 5 0 とを備え、聴覚モジュールが、複数の音響モデルにより音声認識を行ない、各音響モデルによる音声認識結果をセクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【選択図】 図 4

【書類名】 出願人名義変更届（一般承継）
【提出日】 平成15年10月31日
【あて先】 特許庁長官 殿
【事件の表示】
【出願番号】 特願2002-365764
【承継人】
【識別番号】 503360115
【住所又は居所】 埼玉県川口市本町四丁目1番8号
【氏名又は名称】 独立行政法人科学技術振興機構
【代表者】 沖村 憲樹
【連絡先】 〒102-8666 東京都千代田区四番町5-3 独立行政法人科学技術振興機構 知的財産戦略室 佐々木吉正 TEL 03-5214-8486 FAX 03-5214-8417

【提出物件の目録】
【物件名】 権利の承継を証明する書面 1
【援用の表示】 平成15年10月31日付提出の特第許3469156号にかかる一般承継による移転登録申請書に添付のものを援用する。
【物件名】 登記簿謄本 1
【援用の表示】 平成15年10月31日付提出の特第許3469156号にかかる一般承継による移転登録申請書に添付のものを援用する。

特願 2 0 0 2 - 3 6 5 7 6 4

出 願 人 履 歴 情 報

識別番号 [3 9 6 0 2 0 8 0 0]

1. 変更年月日	1 9 9 8 年 2 月 2 4 日
[変更理由]	名称変更
住 所	埼玉県川口市本町 4 丁目 1 番 8 号
氏 名	科学技術振興事業団

特願 2 0 0 2 - 3 6 5 7 6 4

出 願 人 履 歴 情 報

識別番号 [5 0 3 3 6 0 1 1 5]

1. 変更年月日 2 0 0 3 年 1 0 月 1 日
 [変更理由] 新規登録
 住 所 埼玉県川口市本町 4 丁目 1 番 8 号
 氏 名 独立行政法人 科学技術振興機構
2. 変更年月日 2 0 0 4 年 4 月 1 日
 [変更理由] 名称変更
 住 所 埼玉県川口市本町 4 丁目 1 番 8 号
 氏 名 独立行政法人科学技術振興機構
3. 変更年月日 2 0 0 9 年 8 月 2 0 日
 [変更理由] 住所変更
 住 所 東京都千代田区四番町 5 - 3 サイエンスプラザ 5 F
 氏 名 独立行政法人科学技術振興機構
4. 変更年月日 2 0 0 9 年 1 0 月 8 日
 [変更理由] 住所変更
 住 所 埼玉県川口市本町四丁目 1 番 8 号
 氏 名 独立行政法人科学技術振興機構